AD_____

Award Number: DAMD17-96-1-6254

TITLE: Computer-Aided Diagnosis and Feature-Guided Data Reduction Systems in Mammography

PRINCIPAL INVESTIGATOR: Heang-Ping Chan, Ph.D.

CONTRACTING ORGANIZATION: University of Michigan
Ann Arbor, Michigan 48103-1274

REPORT DATE: October 2000

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2000	Annual (23 Sep 99 - 22 Sep 00)

4. TITLE AND SUBTITLE
Computer-Aided Diagnosis and Feature-Guided Data Reduction Systems in Mammography

5. FUNDING NUMBERS
DAMD17-96-1-6254

6. AUTHOR(S)
Heang-Ping Chan, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Ann Arbor, Michigan 48103-1274

E-MAIL:
chanhp@umich.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
This report contains colored photos

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 Words)*

We have been conducting the pilot clinical study to evaluate the effects of CAD on radiologists' reading of screening mammograms this year. We have analyzed the results of about 1,300 cases. The CADView system detected 100% (18/18) of the lesions that were recommended for biopsy in both sites, all fine needle biopsy cases (5/5) at the GU site, and missed only one of the fine needle biopsy cases (4/5) at the UM site. The CAD system detected both malignant cases at the UM site, whereas causing 19 additional callbacks and 1 additional benign biopsy. The CAD system detected all three cancers at the GU site, including one additional cancer that was not originally called by the radiologist, and only caused 2 additional callbacks. The CAD system also detected 74% (34/46) of the short-term follow up cases at the two sites. Since the number of cases collected so far is still small, we have not performed statistical analysis on the data yet. We will continue to collect cases at the UM and GU sites in the coming year.

Two observer performance studies have been conducted for the CAD-guided image compression project. It was found that the proposed method with adequate bit rate will fully preserve the quality of microcalcifications and suspected microcalcifications without sacrificing the edge sharpness and overall image quality at an area-equalized bit-rate of about 0.4 bit/pixel. The CAD-guided compression can therefore reduce the image transmission and storage requirements for digital mammograms by a factor of about 30 without causing observable degradation of image quality. It can be an effective image compression method for picture archiving and communication and facilitate the implementation of telemammography and digital mammography.

14. SUBJECT TERMS
Mammography
Computer-aided diagnosis, breast cancer detection, data compression

15. NUMBER OF PAGES
66

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

(3) Table of Contents

(4) Introduction

We have been developing CAD algorithms in detection of microcalcifications and masses using advanced image processing and computer vision techniques. Our CAD algorithms have provided very promising results in laboratory tests. Our goals in this proposal are to implement our CAD algorithms in a fast workstation, develop user interfaces for efficient operation of the CAD programs, and conduct a pilot clinical trial of the CAD schemes at two mammographic screening sites. Based on the results of the pilot clinical trial, we can evaluate the sensitivity and specificity of the CAD algorithms, analyze the effects of the CAD schemes on mammographic screening, identify any potential problems in a clinical environment, and develop methods to further improve the CAD schemes in the future. We believe that this is a crucial step to develop a clinically practical CAD workstation.

It has been recognized that digital mammography is one of the key research areas for improvement in the diagnosis of breast cancer [1]. Two of the major issues in digital mammography are the technological requirements in developing high resolution digital detectors and the transmission and archiving the large amount of data. Data compression can reduce the amount of data for transmission and storage. However, there is often a tradeoff between compression ratio and image fidelity. Data compression in mammography is especially difficult because of the very subtle image details such as microcalcifications and mass margins that need to be preserved. We have investigated the effects of data compression on computerized detection of microcalcifications previously. In this project, we have developed a CAD guided data compression technique to maximize the compression efficiency with a minimum loss of information. Our approach is to preserve the original image information by lossless compression in potentially important regions on the mammograms indicated by the CAD programs. For breast areas outside these regions, we will apply the most efficient lossy compression technique that does not cause noticeable degradation of image details. We will conduct subjective image quality ranking studies to compare observer performance on the uncompressed images, on images compressed with the selected lossy technique, and on images compressed with the standard JPEG technique.

With the support of this grant from the USAMRMC Breast Cancer Research Program, we have developed a CAD workstation with a proper graphical user interface for a pilot clinical trial. CAD workstations have been implemented at the University of Michigan and at the Georgetown University. In this no-cost-time-extension year, our main goal is to continue to collect cases for the pilot clinical study, and to conduct the observer performance study for comparing compressed and non-compressed mammograms. We will discuss the details of these progresses in the following section.

(5) Body

During the non-cost-time-extension period of 9/23/99 to 9/22/00, our first goal is to conduct the pilot study to collect patient cases in a screening setting. The images will be read by radiologists without and with CAD using the CAD workstations at the University of Michigan and the Georgetown University. Our second goal is to conduct the observer performance to compare compressed and uncompressed mammograms. We have conducted the following tasks:

University of Michigan

(a) CADView workstation

In the previous reports, we have discussed the basic design and operation of our PC-based CAD workstation, "CADView", and its graphical user interface (GUI) in detail. After we conducted training reading sessions with radiologists, we have made further improvement in the GUI, the display, and the data collection system in this year. The current version of the CADView system being used in the pilot clinical study is shown in Figure 1. The reading process is as follows. The radiologist will read the original film mammograms on the alternator as in their daily clinical practice. They will then retrieve the patient 4-view mammogram to be displayed on the CADView monitor by scanning the barcode of the patient folder. They will mark any potential masses on the displayed images and record their impression of the most suspicious mass using the BI-RADS lexicon. They also select the BI-RADS action category for the mass which is recorded by the CAD system. Any potential microcalcification locations will then be marked and the BI-RADS impression and action category for the microcalcifications are recorded. The computer then displays the detected suspicious masses on the images. The radiologist will read the original films again based on the computer prompts. The radiologist can change their initial markings of masses on the displayed images if they are influenced by the computer output. They can also change the BI-RADS impression and the action category for the mass. The same procedure will also be performed for microcalcifications. The markings and action categories of the radiologist before and after CAD display are both recorded in a database file.

Figure 2 illustrates an example of the radiologist's markings on the displayed images. The double circles marked the location of the most suspicious mass in Figure 2(a) and the location of the most suspicious microcalcification clusters in Figure 2(b). The sliders on the right indicated the BI-RADS impression of the marked lesions. The right and left breasts were recorded separately. The BI-RADS action categories for the lesions were also selected on the sliders.

Figure 3 illustrates the same example after the CADView displayed the computer detection output. The computer detected masses were marked by arrowheads and the computer detected clusters were marked by dots. The radiologist's original marks were superimposed on the computer output. If there were disagreements, the radiologist could double-check the film mammograms on the alternator to resolve the discrepancy. If the radiologist found additional suspicious locations, he/she would add new marks on the displayed images. If the new locations were deemed more suspicious than the ones that he/she marked before the computer output was displayed, they could move the double circles to the new locations. The radiologist could also change their BI-RADS impression and action categories on the lesions by moving the pointers on the sliders.

Figure 4 shows the clinical setting of the CADView system. The display is placed next to the offline alternator and the radiologist can easily access the keyboard and mouse. Patient retrieval is through a barcode reader. All other input is through "point and click" by using the mouse. The mammograms displayed on the screen are arranged in exactly the same way as the films mounted on the alternator to facilitate the radiologist to compare the corresponding locations marked on the images.

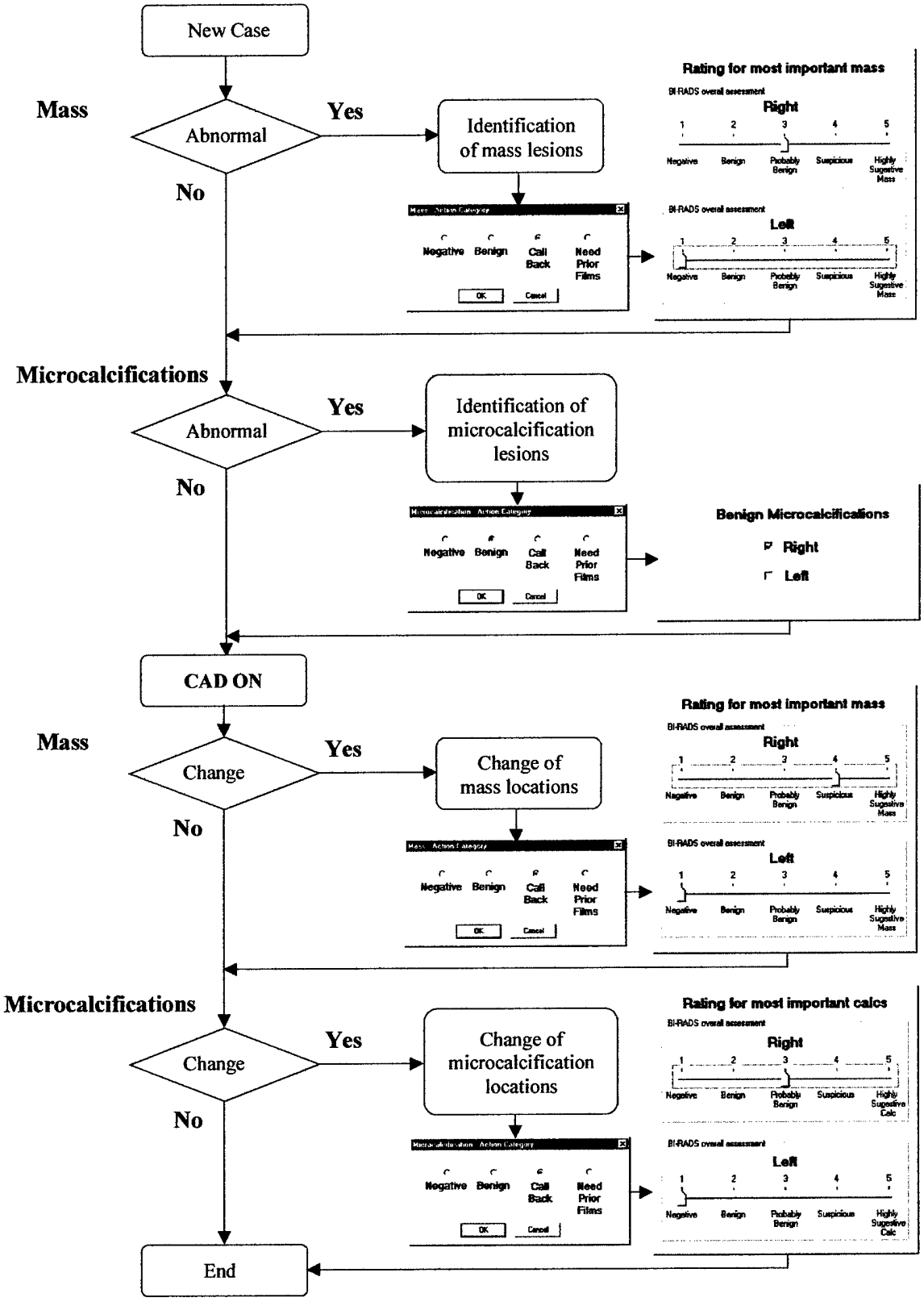


Figure 1. Sequence of reading and collection of the radiologist's BI-RADS assessment of a mammographic case.

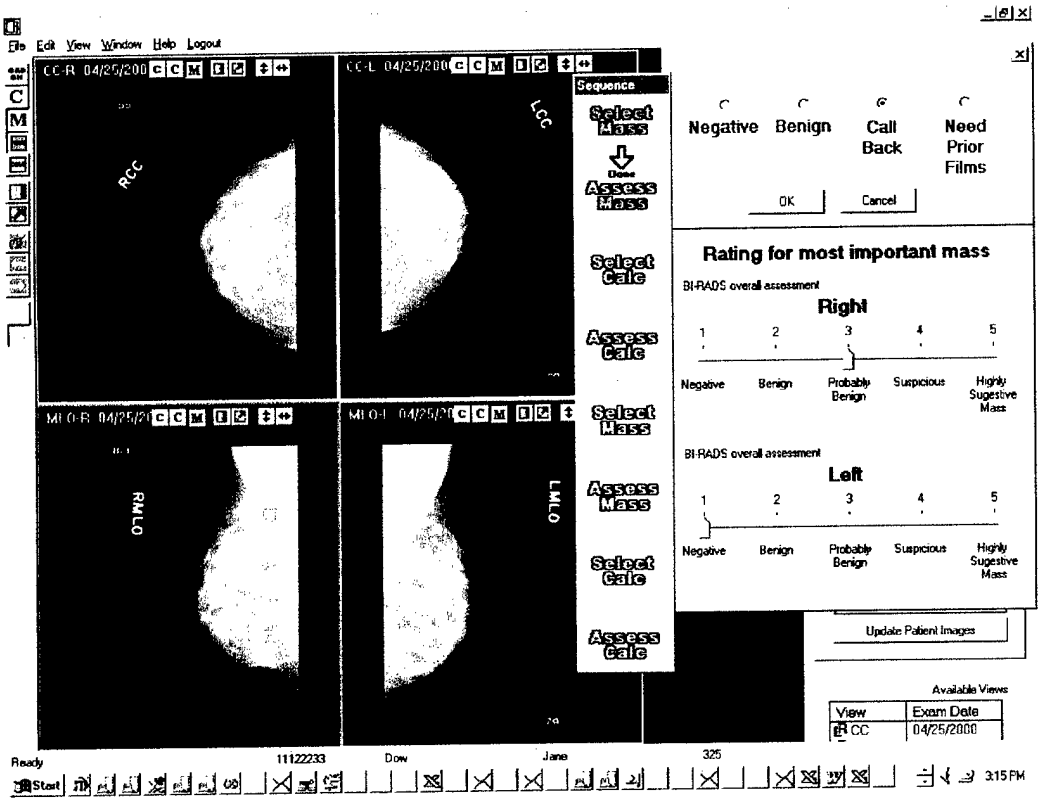


Figure 2(a). Radiologist's assessment of mass before CAD display.

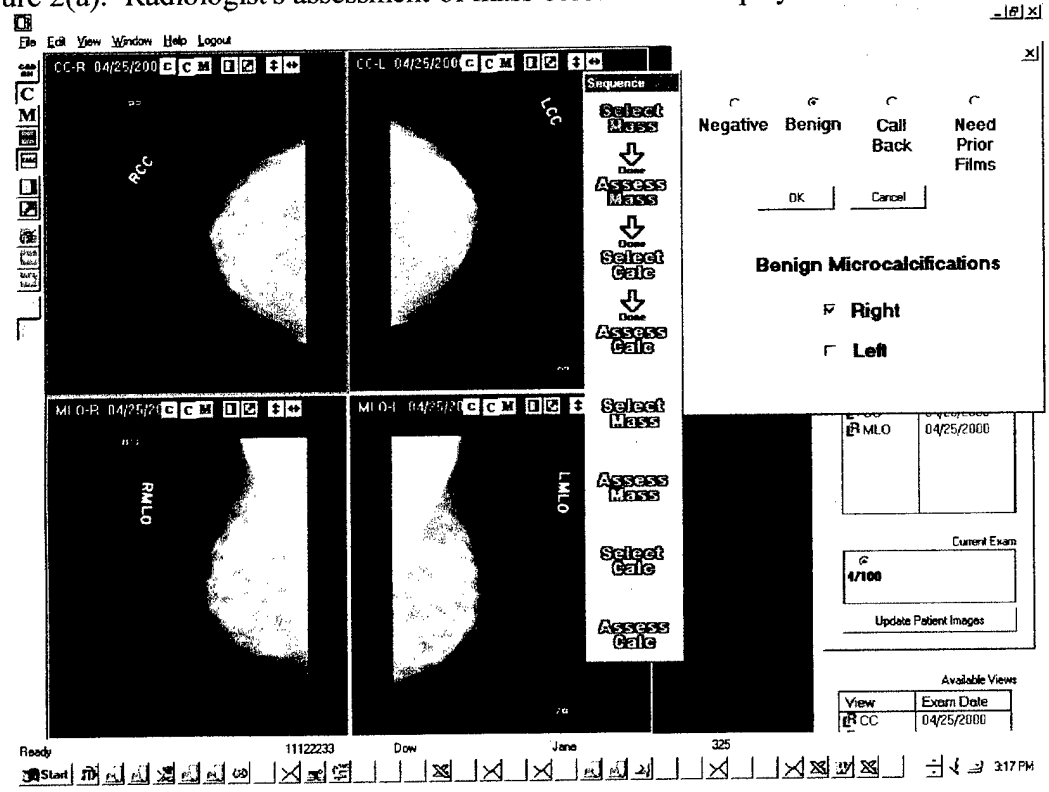


Figure 2(b). Radiologist's assessment of microcalcifications before CAD display.

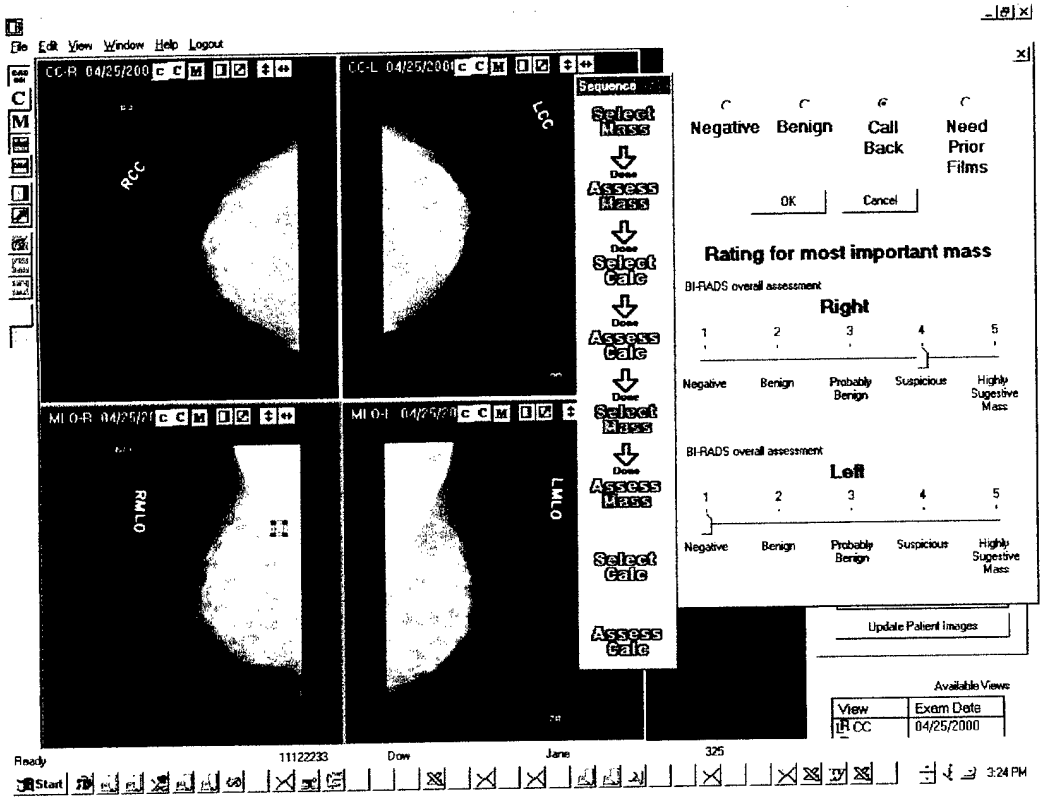


Figure 3(a). Radiologist's assessment of mass with CAD display.

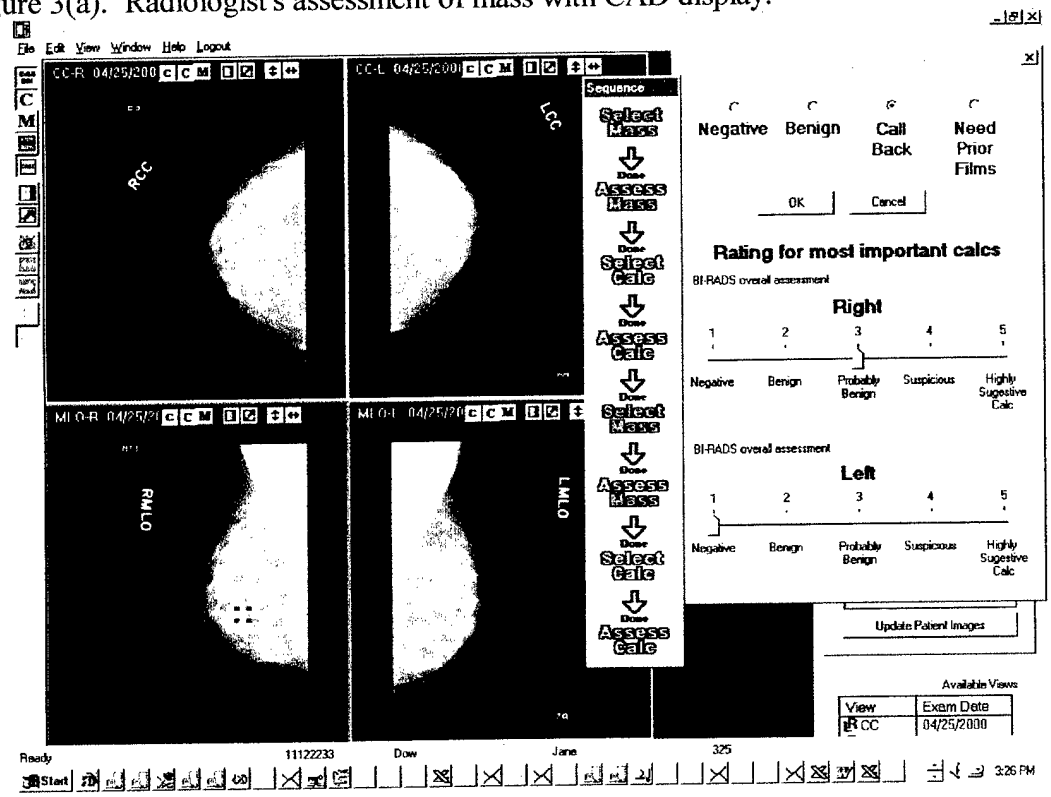


Figure 3(b). Radiologist's assessment of microcalcifications with CAD display.

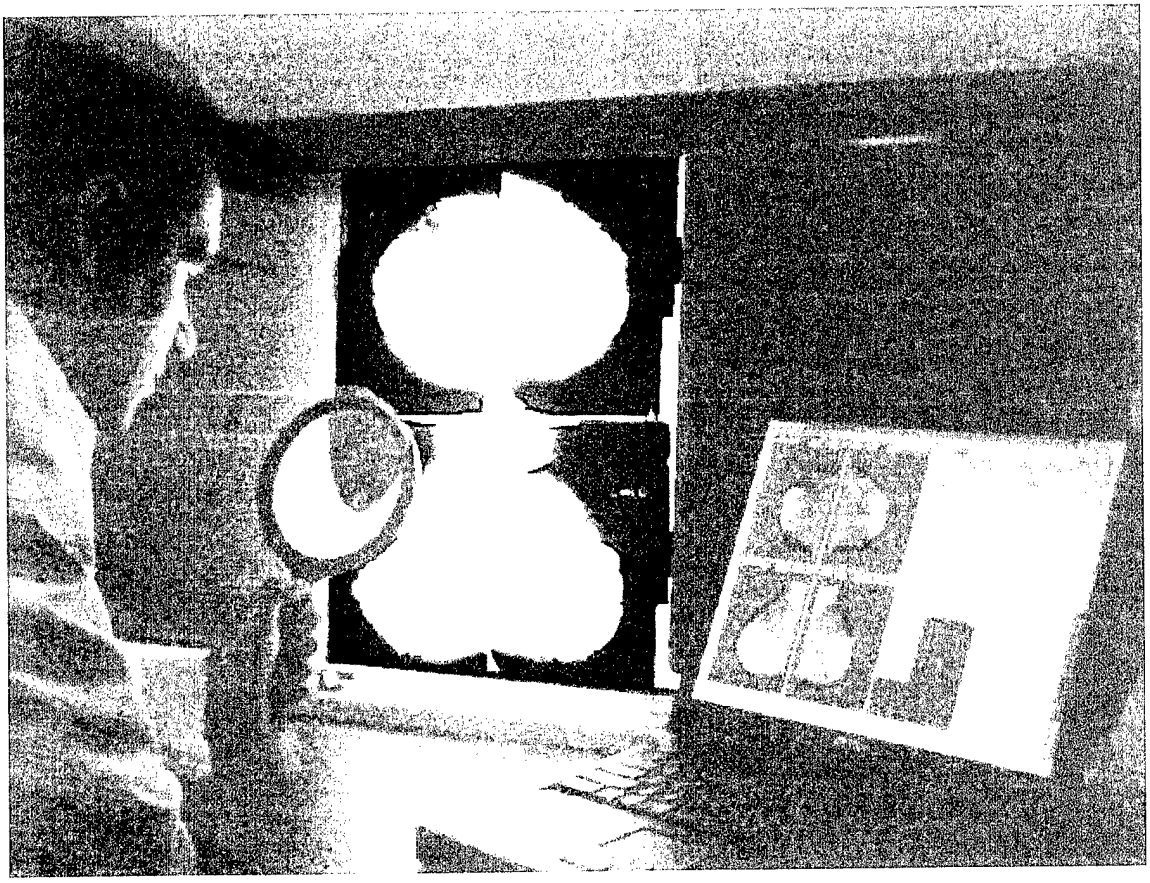


Figure 4. The setup for the CAD reading of off-line screening mammograms. The radiologist reads the original film mammmograms on the alternator while using the computer output as a second opinion.

(b) Collection of screening mammograms

To date, we have collected over 1000 cases at a University of Michigan (UM) breast imaging off-line screening site, and over 400 cases from the Georgetown University (GU) Breast Imaging clinic. At the University of Michigan, the off-line screening cases are transported to the central site at the Breast Imaging Division at the Department of Radiology the next day early in the morning. The films are digitized and processed on the same day. The computer output is ready to be used in the CADView workstation when the screening cases are read the next day. The radiologist will read the cases on the off-line alternator, together with all other screening cases. The radiologists' assessment scores are recorded in the CADView system.

We have analyzed the first 850 cases. We do not have the callback results and follow up information on the other cases yet because of the time delay between a decision to call back and the scheduled call back exam. The number of callbacks, biopsies, and follow-up cases within the first 850 participating patients at the UM are summarized in Table I.

Because the number of callbacks within the patient cohort is still small, we have not performed statistical analysis of the data. From these initial results, we can make some observations:

1. For the cases that the radiologists recommended biopsy, the computer program detected 100% (12/12) of the lesions.
2. For the cases that radiologists recommended fine needle biopsy, the computer program detected 75% (3/4) of the lesions.
3. The computer detected both malignant cases (2/2) found in this patient group.
4. The computer caused 19 additional call backs, of which 6 were recommended 6 month follow-up and 1 was recommended biopsy, indicating that the computer found some areas of concern that the radiologists would not have called without the computer output. The development of the 6-month follow-up cases will be followed.
5. The computer caused 1 additional biopsy that was found to be benign.
6. The computer has a detection sensitivity of 71% for masses and 81% for microcalcifications, similar to our prediction in laboratory tests.
7. The computer missed 1 case that was recommended for fine needle biopsy and found to be benign, and 8 microcalcification or mass cases, that were recommended for 6 month follow up. These 6-month follow-up cases will again be followed.

Table I. The performance of the CADView detection system and its effects on radiologists' reading on the callback cases from the first 850 off-line screening cases at the University of Michigan. The 12 month follow up indicates a regular annual screening schedule and these cases are thus generally considered to be normal.

	Biopsy	Fine needle aspiration	6 month follow up	12 month follow up	Overall	Malignancy
Call Backs for Mass						
Radiologist detection	6	3	20	79	108	1
Computer detection	6	2	14	55	77	1
Sensitivity of computer (%)	100%	67%	70%	70%	71%	100%
Call Backs for Calcs						
Radiologist detection	4		7	5	16	0
Computer detection	4		6	3	13	0
Sensitivity of computer (%)	100%		86%	60%	81%	
Call Backs for Mass and Calcs						
Radiologist detection	2	1	5	5	13	1
Computer detection	2	1	4	5	12	1
Sensitivity of computer (%)	100%	100%	80%	100%	92%	100%
Overall Call Backs						
Radiologist detection	12	4	32	89	137	2
Computer detection	12	3	24	63	102	2
Sensitivity of computer (%)	100%	75%	75%	71%	75%	100%
Call Backs caused by CAD						
Mass	1		2	13	16	
Microcalcifications			2	1	3	
Computer False Negatives						
FN for Calcs			1	2	3	0
FN for Mass		1	6	24	31	0
FN for Mass and Calcs			1		1	0

Georgetown University

Annual Report (9/23/99 – 9/22/00) to USAMRMC through University of Michigan

(a) Continuation of Computer-Aided Detection clinical trial at Georgetown University

Since January 2000, Drs. Freedman and Makariou have used the system to perform a part of their routine clinical readings. We placed a Lumisys film digitizer (Model Lumiscan 150) hosted by a SUN SPARC 10 workstation at the Breast Imaging Division, Radiology Department, Georgetown University Medical Center. A part-time student was hired to enter the patient demographic data and to digitize the films about 2-3 times a week. Since the mammography film loading is completed at 4:30 pm in the afternoon, the student can only work off hour from 5 pm to 9 pm. Usually the student is able to digitize about 20 cases (80 mammograms) for each working session. The data flow is chained through a 3-step procedure.

Step 1: A mammogram is digitized at the SUN/Lumiscan workstation. Patient information, including ID, age, side, view (CC or MLO) and examination date, is recorded during the digitization and entered into CAD patient/film database (part of the CADView system) on the PC computer. Each mammogram is digitized at 100 micron resolution. The image files are stored at a designated directory at the SUN workstation hard disk. The image files are also transferred for further processing to the XP1000 workstation at the ISIS Center via a high-speed Ethernet connection.

Step 2: A control program running on the XP1000 workstation continuously searches for new images being transferred from the SUN/Lumiscan workstation. When a new image appears, this control program initiates the execution of the program to detect the mass and clustered microcalcifications on that image and stores the detection results in appropriate directories.

Step 3: On the PC workstation, the CADView program, designed and implemented by the University of Michigan team, is used as the user interface to review and analyze the results of the mass and microcalcification detection. The CADView program uses an automated procedure to download the output images from the XP1000 workstation on an on-demand basis. The radiologist uses the patient ID number to retrieve patient information from the database (updated in step 2), including the CAD output information on the images to be displayed, and display them on the screen. If the requested images are not available locally, the program establishes an FTP session with the XP1000 workstation and downloads those image files to its working directory on the PC workstation. The radiologist can then perform the clinical evaluation of the patient films. The program, among others, allows the radiologist to mark the location of any suspicious masses and/or microcalcifications on the images, along with his/her action rating. The results of the radiologist's review and evaluation are stored in the database.

(b) Summary of the cases collected at Georgetown University Medical Center

In the past nine months, 1189 cases (4756 images) were digitized and processed by the CAD system. However, only about 40 percent of the cases were reviewed in conjunction with the clinical reading by the radiologists due to some operation issues and mismatch of scheduling. The subcategories of the collected cases are given below.

We have analyzed the first 442 cases from the Georgetown University. The results are analyzed in terms of the callback and follow up decisions, as summarized in Table II. The pathology of some of the biopsy cases is still being tracked. So far only three cases have been identified as malignant. From these initial results, we can make some observations:

1. For the cases that the radiologists recommended biopsy, the computer program detected 100% (6/6) of the lesions.
2. For the cases that radiologists recommended fine needle biopsy, the computer program detected 100% (5/5) of the lesions.
3. The computer detected all three malignant cases (3/3) found in this patient group. Two were mass cases and one was microcalcification case.
4. The computer caused 2 additional call backs, of which 1 was recommended biopsy and found to be malignant.
5. The computer has a detection sensitivity of 75% for masses and 80% for microcalcifications, similar to our prediction in laboratory tests and also similar to the detection sensitivity found at the University of Michigan site. These results confirm that the performance of the CAD system is consistent in the patient population, although the two sites use different digitizers and different mammography systems.
6. The computer missed 4 cases that were recommended for 3 or 6 month short-term follow up. The development of these follow-up cases will be followed.

Table II. The performance of the CADView detection system and its effects on radiologists' reading on the callback cases from the first 442 off-line screening cases at the Georgetown University. The 12 month follow up indicates a regular annual screening schedule and these cases are thus generally considered to be normal.

	Biopsy	Fine needle aspiration	3-6 month follow up	12 month follow up	Overall	Malignancy
Call Backs for Mass						
Radiologist detection	3	5	5	42	55	2
Computer detection	3	5	3	30	41	2
Sensitivity of computer (%)	100%	100%	60%	71%	75%	
Call Backs for Calcs						
Radiologist detection	3		7	10	20	0
Computer detection	3		5	8	16	1
Sensitivity of computer (%)	100%		71%	80%	80%	
Call Backs for Mass and Calcs						
Radiologist detection			2	4	6	
Computer detection			2	3	5	
Sensitivity of computer (%)			100%	75%	83%	
Overall Call Backs						
Radiologist detection	6	5	14	56	81	2
Computer detection	6	5	10	41	62	3
Sensitivity of computer (%)	100%	100%	71%	73%	77%	
Call Backs caused by CAD						
Mass				1	1	
Microcalcifications	1				1	1
Computer False Negatives						
FN for Calcs			2	2	4	
FN for Mass			2	12	14	
FN for Mass and Calcs				1	1	

(c) Image compression of mammograms using a CAD-guided wavelet compression method

Currently, it is possible to obtain a digital mammogram having high spatial resolution by digitizing screen-film images with a laser digitizer [2-4] or a direct digital systems [5,6]. The research and development of teleradiology and telemammography systems have progressed through many technical and clinical endeavors [7-9]. However, the clinical utilization of teleradiology systems is still not known with regards to workloads, reliability, and clinical protocols. The selection of efficient and cost-effective wide-area networks for various applications is presently more an art than a science. In this area, two technical problems remain: (a) no model exists by which radiologists can apply the experience of others to design and implement a teleradiology system; (b) teleradiology systems have not been studied for use in research and education.

(c.1) CAD-guided compression scheme for digital mammography

We randomly selected 100 mammograms from our clinical database. Each of these mammograms contain isolated and/or clustered microcalcifications. The mammograms were digitized by a Lumisys (LumiScan Model 150) at 100 microns pixel size that generates a computer file of 1792x2560x16 bits. However, only 12 out of 16 bits were used to store the digital data for each pixel.

Prior to performing the wavelet transform, the boundary of the mammogram was delineated. Only the area within the boundary was compressed. We used an integer wavelet transform [10,11] to decompose the mammogram followed by a linear quantization process and arithmetic coding to encode the quantized wavelet coefficients. In order to preserve the data accuracy of calcifications and suspected calcifications, we employed our computer-aided detection procedure that can detect excessive number of small bright spots on the mammogram. All the suspected calcifications and their adjacent areas were then losslessly encoded. The major reason to apply lossless coding on all suspected calcifications is two-fold: (1) to preserve the original quality of calcifications which are clinically significant features associated with breast cancer, and (2) to maintain the original quality of calcification-like spots that may otherwise become false-positives due to the blurry effect of the compression. Figure 5 illustrates the compression scheme used in this study.

We decomposed each image with 5-level wavelet transform; hence, the matrix size of the smallest image was 112x160 pixels. The lowest resolution subimage was further decomposed by an error-free compression method (i.e., A DPCM followed by an arithmetic coding). The bit-allocation and quantization were determined based on the energy concentration in each level of the high frequency domains. Beside the quantization, all data processing procedures are reversible.

The decompression was done by inverse arithmetic coding to resume the quantized coefficients on the wavelet domain followed by an inverse wavelet transform. The inverse transformed image is a compressed version of the mammogram that possesses small variances on the majority of the pixels. The compressed data on the B file was processed by another inverse arithmetic coding process. The reconstructed data was added onto the pixel values of the suspected areas.

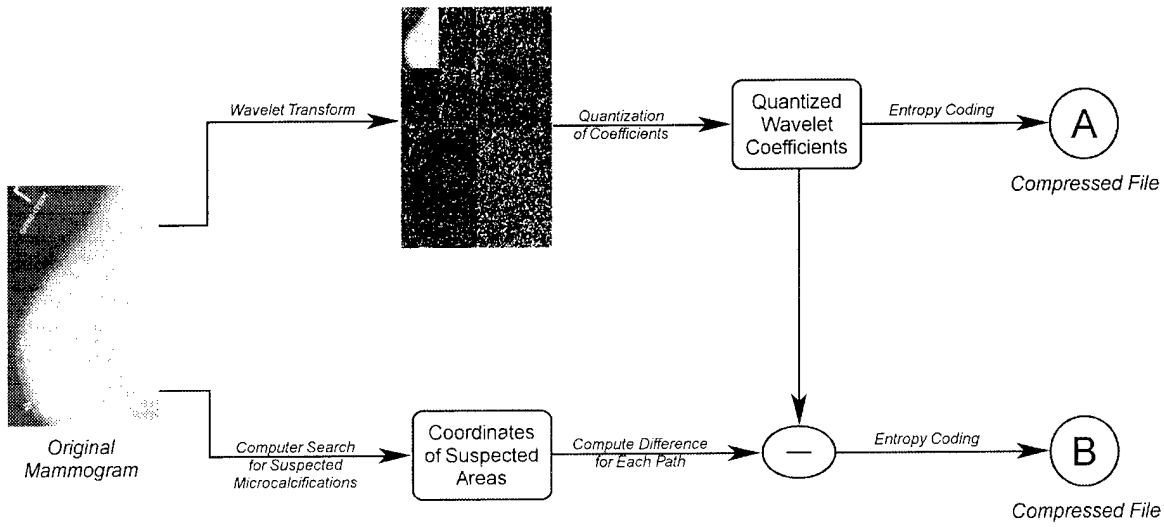


Figure 5: A CAD guided compression scheme based on integer wavelet decomposition.

(c.2) Description of observer performance studies

We asked a senior breast radiologist to view a hundred sets of images with four different compression modes and to rate their impressions of their comparative quality. Each set of images is a pair of original and one of three compression modes. The three compression modes are: (i) 0.3 bit/pixel date wavelet encoded in compressed file A with the residual data for lossless compression of suspected calcifications in file B, (ii) 0.1 bit/pixel wavelet encoded in compressed in file A with the residual data for lossless compression of suspected calcifications in file B, and (iii) 0.1 bit/pixel data wavelet encoded in file A only.

Each set of decompressed and original images were randomly displayed on two SUN computer monitors (right or left) as a pair. The effective image size was approximately magnified by a factor of 4 (i.e., 2x2). Contrast and brightness controls were available as software functions for the radiologist to adjust the viewing parameters when necessary. A synchronized display software was developed for the comparative visual study. The software allows the user to simultaneously display and control image functions on the paired images using a single or two monitors. The reader was asked to rate image quality in terms of calcification observability, edge sharpness, overall image quality and to rate noise appearance for all images. A four-section questionnaire was used and is shown in Figure 6.

Letters "L" and "R" indicate that the left or right sides rank higher on the dimension measured, respectively. A non-zero score indicates that one side of the image has either slightly (for L1 or R1), or moderately (for L2 or R2), or significantly (for L3 or R3) better quality or more noise than the other side. A score of "0" indicates that the pair of images has identical image quality or noise appearance. If there is some noticeable difference between images that are scored "0" on the measured dimension, this is indicated by checking the bottom box below the "0" score. If reader is in favor of one image for its specific feature, one of the two boxes (left and right) can be checked to indicate his/her preference.

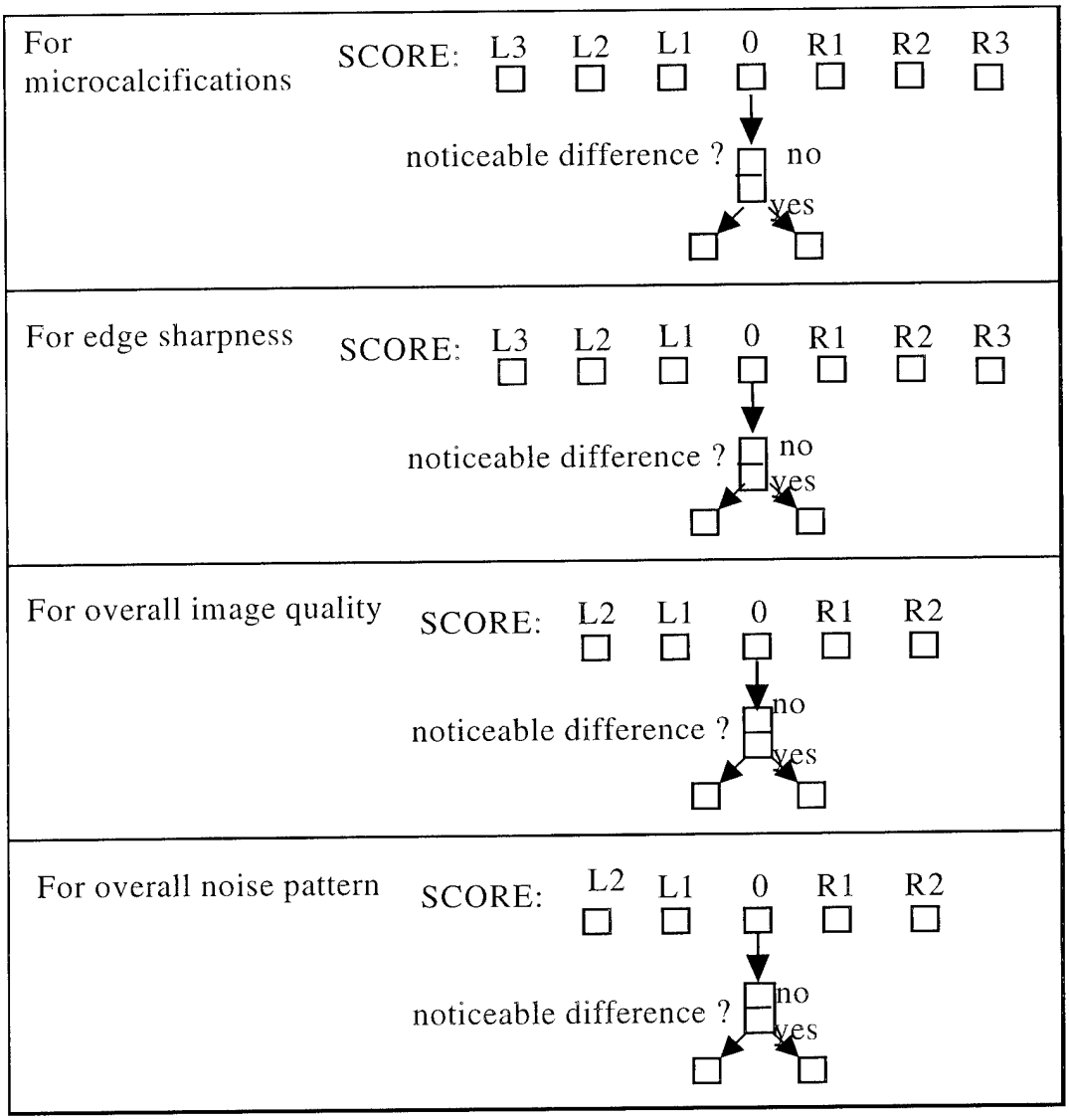


Figure 6: A questionnaire for qualitative measures for a pair of images.

(c.3) Observer experiments and results

(1). The First Observer Study

The modified CAD program detected an average of 1,193 potential microcalcifications in CC view mammograms and an average of 948 potential microcalcifications in MLO view mammograms, respectively. Figures 7 and 8 show two sample mammograms, their compressed counterparts, and the subtracted images.

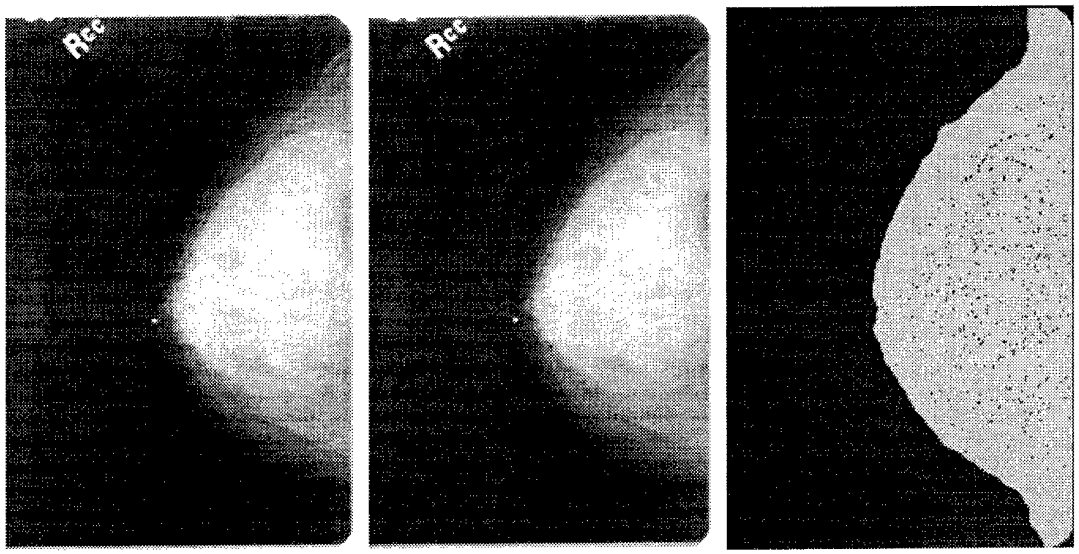


Figure 7: (A) A CC view mammogram, (B) its compressed image at 0.4 bit/pixel, and (C) the enhanced subtracted image resulting from (A)-(B). The uniform squares in (C) result form the lossless compression at the CAD detected areas.

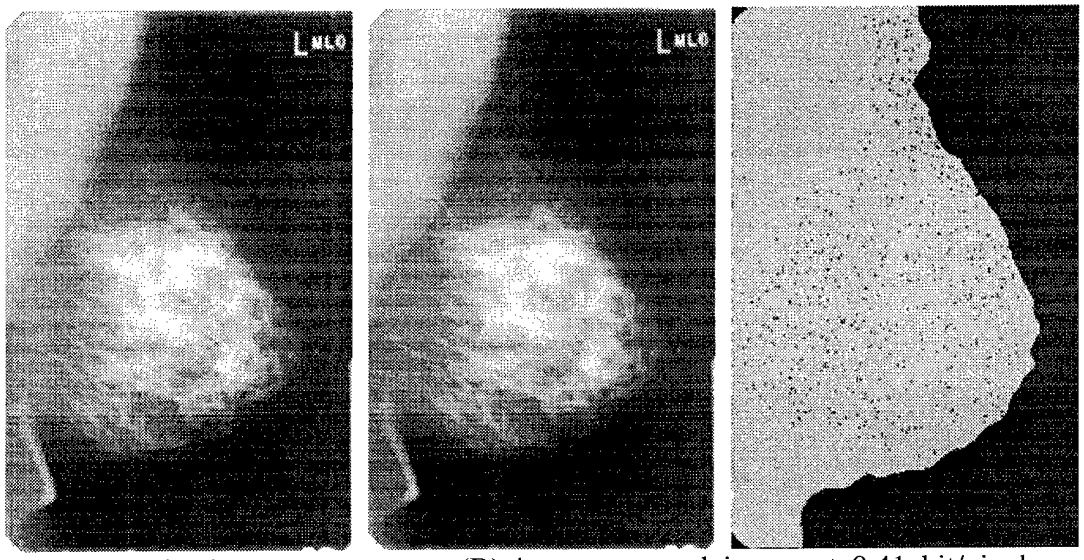


Figure 8: (A) A MLO view mammogram, (B) its compressed image at 0.41 bit/pixel, and (C) the enhanced subtracted image resulting from (A)-(B). The uniform squares in (C) result form the lossless compression at the CAD detected areas.

The average compression ratios and computed mean-square-errors (MSE) between the original and decompression are shown in Table III. We found that the CAD guided compression method received very small MSE improvement although it used a significant number of computer space (or bit rate) to preserve the full data accuracy of the suspected calcifications. This mainly is because that the suspected microcalcifications occupy very small area as compared to the whole breast region.

Table III. Compression Ratios and Mean-Square-Errors of the Three Compression Modes in the First Observer Study.

Mode	A	B	C
Procedure	0.3 bit/pixel + lossless for spots	0.1 bit/pixel + lossless for spots	0.1 bit/pixel
Average Bit Rate	0.43 bit/pixel	0.23 bit/pixel	0.1 bit/pixel
Compression Ratio	27:1	52:1	120:1
Mean Square Error (Standard Devation)	50.73 (36.81)	102.72 (62.48)	105.63 (63.97)

Table IV. Qualitative measures by comparing the paired images in the First Observer Study (Original and Compressed).

Measurement Category	Micro-Calcifications			Edge Sharpness			Overall Image Quality			Overall Noise Pattern		
	Type A	Type B	Type C	Type A	Type B	Type C	Type A	Type B	Type C	Type A	Type B	Type C
Original Worse Than Compressed	7	4	1	6	2	3	1	0	2	1	0	0
of which: - same, but in favor of compressed	3	4	1	5	2	0	0	0	2	1	0	0
- slightly worse	4	0	0	1	0	3	1	0	0	0	0	0
- moderately worse	0	0	0	0	0	0	0	0	0	0	0	0
Original Better Than Compressed	7	5	10	4	11	15	1	7	7	0	2	3
of which: - same, but in favor of original	6	1	4	4	6	9	1	3	2	0	2	0
- slightly better	1	4	5	0	5	5	0	4	4	0	0	3
- moderately better	0	0	1	0	0	1	0	0	1	0	0	0
No Difference	36	16	14	40	12	7	48	18	16	49	23	22

Type A - Compression with preservation of suspicious calcifications; Compression rate: 0.43 bit/pixel (0.3+0.13); Total 50 Cases
Type B - Compression with preservation of suspicious calcifications; Compression rate: 0.23 bit/pixel (0.1+0.13); Total 25 Cases
Type C - Global compression; Compression rate: 0.1 bit/pixel; Total 25 Cases

Table IV illustrates the results of the radiologist's qualitative measures while comparing the original and compressed image pair. We found that no difference could be observed between the original and decompressed images at a bit rate of 0.43 bit/pixel. In fact, it is interesting that the radiologist seemed slightly in favor of the appearances of microcalcifications and edges in the compressed mammograms. The radiologist identified 20% of the compressed images at 0.1 bit rate suffering from minor blurring artifacts and 6% of the compressed images possessing greater edge sharpness. Without using lossless compression for microcalcifications, the radiologist could identify

20% of the less sharp microcalcifications on the compressed mammograms at 0.1 bit rate. The radiologist also identified that 18% and 6% of the compressed images at 0.1 bit rate possess degraded overall image quality and higher image noise, respectively. Degradation of image quality in compressed images at 0.1 bit rate is highly associated with unsharpness of microcalcifications and edges. The image quality degradation at 0.1 bit rate is also correlated with the size of breast area. It is estimated that if the size of the breast takes more than one half of the entire mammogram, degradation in image quality and edge unsharpness would be observed by the radiologist.

We also compared the compression rate and breast area and successfully identified cases that possessing higher breast area would suffer observable low quality. On the other hand, for compression rates higher than or equal to 0.1bit/pixel and breast area less than or equal to 40%, no degradation can be identified relatve to their original counterpart in overall image quality, overall noise pattern, and edge sharpness. For compression rates higher than or equal to 0.1bit/pixel and breast area less than or equal to 25%, no degradation can be identified as inferior microcalcifications. Therefore, we pre-assessed that the threshold of area-equalized compression rate for the background including edges is 0.25 bit/pixel (0.1bit/pixel divided by 40%) and the threshold of area-equalized compression rate for the microcalcification is approximately 0.4 bit/pixel (0.1 bit/pixel divided by 25%).

(2). The Second Observer Study

In this study, we compared two different compression methods: (1) using an area-equalized compression rate at 0.25 bit/pixel with preservation of microcalcifications to compress and decompress the mammograms and (2) using an area-equalized compression rate at 0.4 bit/pixel to compress and decompress the mammograms. The conventional bit rate and area-equalized (AEQ) bit rate are defined below:

Bit rate = total number of bits used to encode the data / total number of pixels in the image (1)

AEQ bit rate = total number of bits used to encode the data / total number of pixels within the breast. (2)

The two decompressed images from the same original mammograms using the two compression methods were randomly displayed on the monitors (right or left). All one hundred cases were used in this study. Other reading parameters and setting were identical as in the first experiment. The reader was asked to rate image quality in the same four image quality categories. The same questionnaire was used.

In this experiment, no image was rated better than its counterpart by the radiologists. However, the radiologist favored microcalcifications of 55 cases that were compressed and decompressed through the first method (i.e., 0.25 AEQ bit/pixel with preservation of microcalcifications). However, the radiologist also favored edge characteristics of 8 cases that were compressed and decompressed through the first method. We believe that the radiologist's assessments in these eight cases were somewhat influenced by favoring the microcalcifications on the images. No image was identified as a higher quality image over its counterpart by the radiologist in terms of overall image quality and overall noise pattern. No image compressed by the second compression method (i.e., 0.4 AEQ bit/pixel) was in favor by the radiologists. Table V shows the summary results of the observer study. Table VI shows the bit rate used and the average MSE of the decompression images for each category. Note that the bit rate of the first method includes the wavelet compressed data and the lossless compressed data of the suspected

calcification areas. Although the first compression method spent less computer space to code the overall breast area than the second method did, the first compression method used more computer space to preserve the 10x10 pixels area of all suspected microcalfications, the effective compression bit rates were approximately the same for both methods. We found that the first method produced higher quality in clinically significant features. Although the overall MSEs produced by the first compression method were markedly worse than those produced by the second method, the degradation was not observable by the breast radiologist. Nevertheless, the first compression method generates error-free suspected calcifications that were appreciable and in favor by the radiologist.

Table V. Qualitative measures by comparing the paired images in the Second Observer Study. (Compression Methods 1 and 2).

Category	Micro-calcifications	Edge Sharpness	Overall Image Quality	Overall Noise Pattern
Total	100	100	100	100
of which:				
In favor of the first method	55	8	0	0
In favor of the second method	0	0	0	0
No Difference	45	92	100	100

Table VI. Compression Ratios and Mean-Square-Errors of the Two Compression Methods in the Second Observer Study.

Category	The First Method (0.25 AEQ bit/pixel + Lossless spots)		The Second Method (0.40 AEQ bit/pixel)	
	Bit Rate (Bit/pixel) Mean (SD)	MSE	Bit Rate (Bit/pixel) Mean (SD)	MSE
All	0.149(0.05)	94	0.141(0.05)	55
Micro-calcifications:				
In favor of the first method	0.146(0.06)	94	0.135(0.05)	65
No Difference	0.152(0.04)	93	0.148(0.05)	53
Edge Sharpness:				
In favor of the first method	0.195(0.09)	92	0.159(0.08)	49
No Difference	0.145(0.05)	95	0.140(0.05)	55

(c.4) Conclusions and discussion of the compression studies

In this study, we used conventional compression testing methods with and without the CAD guidance to evaluate the decompressed images. We were able to identify the threshold of area-equalized bit rate for overall breast area and the threshold for encoding quality microcalcifications. We used these two thresholds to compress the mammograms. All four image-quality categories of all compression images were deemed more than adequate. However, the radiologist favored fully preserved

microcalcifications on 55 out of 100 images (55% of the test database). This study also showed that neither edge nor overall image quality degradation could be observed by the radiologist using area-equalized bit-rate of 0.25 AEQ bit/pixel and 0.4 AEQ bit/pixel. Therefore, CAD can be used to guide image processing method to preserve or enhance clinically significant features. Our results clearly indicate that the CAD guided compression method with adequate bit rate will fully preserve the quality of microcalcifications and suspected microcalcifications without sacrificing the edge sharpness and overall image quality.

Approximately 11,500 suspected calcifications were reconstructed in this study. These CAD detected suspected areas were losslessly decompressed at their original places among other breast parenchyma on the lossy compressed mammograms. The radiologist could not recognize any blocky artifact between lossless and lossy boundaries even on magnified view with contrast adjustable display. These differences, however, are observable on the enhanced subtraction images as shown in Figures 7 and 8.

In this project, we proposed to use a highly sensitive CAD system to guide the compression method to preserve clinically significant image patterns. Our study demonstrates that the success of this approach.

(6) Key Research Accomplishments

- Start the pilot clinical study at the Breast Imaging clinic of University of Michigan Health System and at the Georgetown University Medical Center
- Collect about 1,400 patient cases with and without CAD reading at the two sites.
- Analyze the effects of CAD on radiologists' reading on 1,300 cases and estimate the performance of the CAD system in the patient population.
- Conduct two observer performance studies to compare microcalcification detection on mammograms without compression, with conventional compression, and with CAD-guided compression

(7) Reportable Outcomes

Publications

University of Michigan

Journal Articles

1. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Medical Physics 1999; 26: 2654-2668.

2. Sanjay-Gopal S, Chan HP, Wilson TE, Helvie MA, Petrick N, Sahiner B. A regional registration technique for automated interval change analysis of breast lesions on mammograms. Medical Physics 1999; 26: 2669-2679.

3. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Classification of malignant and benign masses based on hybrid ART2LDA approach. IEEE Transactions on Medical Imaging 1999; 18: 1178-1187.

Articles Accepted for Publication:

1. Chan HP, Helvie MA, Petrick N, Sahiner B, Adler DD, Paramagul C, Roubidoux MA, Blane CE, Joynt LK, Wilson TE, Hadjiishi LM, Goodsitt MM. Digital mammogramphy: observer performance study of effects of pixel size on radiologists' characterization of malignant and benign microcalcifications. Academic Radiology.

Articles Submitted for Publication:

1. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA. Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis – local affine transformation for improved localization. Medical Physics.

2. Zhou C, Chan HP, Petrick N, Helvie MA, Goodsitt MM, Sahiner B, Hadjiiski LM. Computerized image analysis: Estimation of breast density on mammograms. Medical Physics.

Book Chapter:

1. Chan HP, Petrick N, Sahiner B. Chapter 6. Computer-aided breast cancer diagnosis. In: Artificial Intelligence Techniques in Breast Cancer Diagnosis and Prognosis. Pp. 179-264. Ed. Jain A, Jain A, Jain S, Jain LC, *Series in Machine Perception and Artificial Intelligence*, Vol. 39 (World Scientific: NJ), 2000.

Conference Proceedings

1. Petrick N, Chan HP, Sahiner B, Helvie MA, Paquerault S. Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms. Proc. SPIE 3979. 2000: 967-973.

2. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Paquerault S, Zhou C. Interval change analysis in temporal pairs of mammograms using a local affine transformation. Proc. SPIE 3979. 2000: 847-853.

3. Petrick N, Sahiner B, Chan HP, Helvie, MA, Paquerault S. Preclinical evaluation of a CAD algorithm for early detection of breast cancer. In: Proceedings of The 5th International Workshop on Digital Mammography. IWDM-2000. Toronto, Canada. June 11-14, 2000 (in press).

Abstracts and Presentations

1. Petrick N, Chan HP, Sahiner B, Helvie MA, Paquerault S. Evaluation of an automated computer-aided diagnosis system for the detection of masses on prior mammograms. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000.

2. Hadjiiski LM, Chan HP, Sahiner B, Petrick N, Helvie MA, Paquerault S, Zhou C. Interval change analysis in temporal pairs of mammograms using a local affine transformation. Presented at the SPIE International Symposium on Medical Imaging, San Diego, CA, February 12-18, 2000.

3. Petrick N, Sahiner B, Chan HP, Helvie, MA, Paquerault S. Preclinical evaluation of a CAD algorithm for early detection of breast cancer. Presented at The 5th International Workshop on Digital Mammography. IWDM-2000. Toronto, Canada. June 11-14, 2000.

4. Chan HP, Hadjiiski L, Petrick N, Helvie MA, Sahiner B, Paramagul C, Gurcan MN, Lo SCB., Freedman MT, Dorfman DD, Berbaum KS. Pilot clinical study of a computer-aided diagnosis workstation for mammography. Presented at the Era of Hope Meeting, U. S. Army Medical Research and Materiel Command, Department of Defense, Breast Cancer Research Program, Atlanta, Georgia, June 8-12, 2000.

5. Hadjiiski L, Petrick N, Chan HP, Sahiner B, Helvie MA, Zhou C, Gurcan MN, Paquerault S. Regional registration of masses on current and prior mammograms using DWCE segmentation. Presented at the Chicago 2000-World Congress on Medical Physics and Biomedical Engineering. Chicago, Illinois, July 23-28, 2000.

6. Zhou C, Chan HP, Petrick N, Goodsitt MM, Paramagul C, Hadjiiski LM. Computerized image analysis: breast segmentation and nipple identification on mammograms. Presented at the Chicago 2000-World Congress on Medical Physics and Biomedical Engineering. Chicago, Illinois, July 23-28, 2000.

7. Chan HP, Sahiner B, Hadjiiski LM, Petrick N, Helvie MA, Goodsitt MM. Computer-aided breast cancer diagnosis: Effects of pixel size on computerized classification of microcalcifications in comparison with radiologists' performance. Accepted for presentation at

the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 26-Dec. 1, 2000, Chicago, Illinois.

8. Zhou C, Chan HP, Helvie MA, Petrick N, Goodsitt MM, Sahiner B, Hadjiiski LM. Computer-aided estimation of mammographic breast density. Accepted for presentation at the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 26-Dec. 1, 2000, Chicago, Illinois.

9. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan MN. Computer-aided classification of malignant and benign breast masses by analysis of interval change of features in temporal pairs of mammograms. Accepted for presentation at the 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Nov. 26-Dec. 1, 2000, Chicago, Illinois.

10. Sahiner B, Petrick N, Chan HP, Paquerault S, Helvie MA, Hadjiiski LM. Recognition of lesion correspondence on two mammographic views - a new method of false-positive reduction for computerized mass detection. Accepted for presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2001.

11. Paquerault S, Petrick N, Chan HP, Sahiner B, Dolney AY. Improvement of mammographic lesion detection by fusion of information from different views. Accepted for presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2001.

12. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA, Gurcan M. Analysis of temporal change of mammographic features for computer-aided characterization of malignant and benign masses. Accepted for presentation at the SPIE International Symposium on Medical Imaging, San Diego, CA, February, 2001.

13. Paquerault S, Petrick N, Chan HP, Sahiner B, Helvie MA. Computer-Aided Breast Cancer Diagnosis: Fusion of Information from Two Mammographic Views. Accepted for presentation at the Univ. of Michigan Cancer Research Symposium, December 8, 2000. Ann Arbor, Michigan.

Georgetown University

1. Li H, Wang Y, Liu KJR, Lo SB, and Freedman MT, "Statistical Model Supported Approach to Radiographic Mass Detection -- Part I: Improving Lesion Characterization by Morphological Filtering and Site Segmentation," to appear IEEE Trans. on Medical Imaging, 2000.

2. Li H, Wang Y, Liu KJR, Lo SB, and Freedman MT, "Statistical Model Supported Approach to Radiographic Mass Detection -- Part II: Decision Making through Modular Neural Networks and Hierarchical Visual Explanation," to appear IEEE Trans. on Medical Imaging, 2000.

3. Lo SB, Delegacz A, Freedman MT, Chan HP, Dorfman DD, and Berbarn K.,"Evaluation of Digitized Mammograms Compressed by an Optimized Wavelet Technique and Computer-Aided System," Presented at the Era of Hope Meeting, U. S. Army Medical Research and Materiel Command, Department of Defense, Breast Cancer Research Program, Atlanta, Georgia, June 8-12, 2000.

4. Lo SB, Makariou E, Delegacz A, Chan HP, Dorfman DD, Freedman MT, and Berbarn K. "Integer Wavelet Compression Guided by a Computer-Aided Detection System in Mammography", SPIE Med. Imging, 2001, (in press).

5. Lo SB and Zhao H, "A Filter for Global Region Segmentation: Discoveries of Image Filters through Convolution Neural Network Training," Submitted to IEEE Trans. on Medical Imaging.

(8) Conclusions

We have been conducting the pilot clinical study of the effects of CAD on radiologists' reading of screening mammograms this year. We have collected over 1,400 cases and analyzed the results of about 1,300 cases. The overall sensitivity of the CADView system was found to be reasonably close to our prediction based on laboratory tests and also is consistent between the two sites. More importantly, the computer detected 100% of the lesions that were recommended for biopsy in both sites, all fine needle biopsy cases at the GU site, and missed only one of the fine needle biopsy cases at the UM site. 71% and 75% of the short-term follow up cases were detected by the CAD system. Whether any of the missed short-term follow up cases will turn out to be malignant remains to be followed. The CAD system caused 19 additional callbacks at the UM site, of which 6 were recommended short-term follow up. We will also track these follow-up cases to determine if any of them will turn out to be malignant. The CAD system only caused 2 additional callbacks at the GU site and one of these was found to be malignant. The CAD system detected both malignant cases at the UM site, whereas causing one additional benign biopsy. It detected one additional cancer at the GU site that was not originally called by the radiologist and two other cancers that were also detected by radiologists. The pathology of some other cases at the GU site is still being tracked. Since the number of cases collected so far is still small our collaborator at the University of Iowa was not able to perform statistical analysis on the data yet. We will continue to collect cases at the UM and GU sites in the coming year.

Since the cancer rate in the screening population is only 3 to 5 per 1000, the number of patients planned for this pilot clinical study will not be sufficient to draw statistically significant conclusion on the effects of CAD on the sensitivity of mammographic screening. However, this pilot study will provide an evaluation of the performance of the CAD system in the clinical screening environment and, more importantly, an assessment of the effects of CAD on the callback rate of the radiologists for reading screening mammograms. The results obtained from this pilot study will be important for the design of a large-scale pivotal clinical study in the future.

Two observer performance studies have been conducted for the CAD-guided image compression project. It was found that the CAD guided compression method with adequate bit rate will fully preserve the quality of microcalcifications and suspected microcalcifications without sacrificing the edge sharpness and overall image quality. Neither edge nor overall image quality degradation could be observed by the radiologist using area-equalized bit-rate of 0.25 bit/pixel and 0.4 bit/pixel. The CAD-guided compression can therefore reduce the image transmission and storage requirements for digital mammograms by a factor of 30 to 50 without causing perceivable degradation of image quality. An effective image compression method for picture archiving and communication will facilitate the implementation of telemammography and digital mammography. Both approaches are expected to improve patient care, especially in remote and rural areas.

Because of the budget reduction, the change in the strategy for the CAD workstation development, and the addition of the mass detection program, as described in the previous reports, as well as the incompatibility of different workstations and operating systems, there was a delay in starting the pilot clinical study. We have requested and obtained approval for a no-cost-time-extension of another year to continue collecting patient cases.

(9) References

1. Shtern F, Stelling C, Goldberg B and Hawkins R, "Novel technologies in breast imaging: National Cancer Institute perspective," Society of Breast Imaging, Orlando, Florida, 153-156 (1995).

2. Alvarez RE, in: Recent Developments in Digital Imaging. Eds. Doi K, Lanzl L, Lin PJP. AAPM

3. Brody WR, Digital Radiography. (Raven Press, NY, 1984).

4. Lo SC, Butson P, Lin JS, Hasegawa A, and Mun SK, "Performance Characteristics of Ultra High-Resolution CCD Film Scanners" SPIE Proc. Med. Imaging 1995, (to be published)

5. Sonoda M, Takano M, Miyahara J, and Kato H, "Computed Radiography Utilizing Scanning Laser Stimulated Luminance", Radiology, 1983; 148: 833-838.

6. Ishida M, Kato H, Doi K, Frank PH, "Development of a New Digital Radiographic Image Processing System", Proc SPIE 347;1982:42.

7. Steckel RJ, "Daily X-Ray Rounds in a Large Teaching Hospital Using High-Resolution Closed-Circuit Television," Radiology 1972;105:321.

8. Jelaso DV, Southwonh G, Purcell LH, "Telephone Transmission of Radiographic Images," Radiology 1978; 127:147-149.

9. Kagetso NJ, Zulauf DRP, Ablow RC, "Clinical Trial of Digital Teleradiology in the Practice of Emergency Room Radiology," Radiology 1987;165:551-554.

10. Said A, and Pearlman WA, "A New, Fast, and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees," IEEE Trans. On Circuits and Systems for Video Technology, vol. 6, pp. 243-250. 1996.

11. Lo SC, Xuan J, Li H, Wang Y, Freedman MT, and Mun SK, "Dyadic Decomposition: A Unified Perspective on Predictive, Subband, and Wavelet Transforms", SPIE Proceedings on Medical Imaging, 1997, vol. 3031, pp. 286-301.

(10) Appendix

Publications enclosed

Journal Articles

1. Chan HP, Sahiner B, Wagner RF, Petrick N. Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers. Medical Physics 1999; 26: 2654-2668.

2. Sanjay-Gopal S, Chan HP, Wilson TE, Helvie MA, Petrick N, Sahiner B. A regional registration technique for automated interval change analysis of breast lesions on mammograms. Medical Physics 1999; 26: 2669-2679.

3. Hadjiiski LM, Sahiner B, Chan HP, Petrick N, Helvie MA. Classification of malignant and benign masses based on hybrid ART2LDA approach. IEEE Transactions on Medical Imaging 1999; 18: 1178-1187.

A regional registration technique for automated interval change analysis of breast lesions on mammograms

S. Sanjay-Gopal, Heang-Ping Chan,[a] Todd Wilson, Mark Helvie, Nicholas Petrick, and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

(Received 11 November 1998; accepted for publication 13 September 1999)

Analysis of interval change is a useful technique for detection of abnormalities in mammographic interpretation. Interval change analysis is routinely used by radiologists and its importance is well-established in clinical practice. As a first step to develop a computerized method for interval change analysis on mammograms, we are developing an automated regional registration technique to identify corresponding lesions on temporal pairs of mammograms. In this technique, the breast is first segmented from the background on the current and previous mammograms. The breast edges are then aligned using a global alignment procedure based on the mutual information between the breast regions in the two images. Using the nipple location and the breast centroid estimated independently on both mammograms, a polar coordinate system is defined for each image. The polar coordinate of the centroid of a lesion detected on the most recent mammogram is used to obtain an initial estimate of its location on the previous mammogram and to define a fan-shaped search region. A search for a matching structure to the lesion is then performed in the fan-shaped region on the previous mammogram to obtain a final estimate of its location. In this study, a quantitative evaluation of registration accuracy has been performed with a data set of 74 temporal pairs of mammograms and ground-truth correspondence information provided by an experienced radiologist. The most recent mammogram of each temporal pair exhibited a biopsy-proven mass. We have investigated the usefulness of correlation and mutual information as search criteria for determining corresponding regions on mammograms for the biopsy-proven masses. In 85% of the cases (63/74 temporal pairs) the region on the previous mammogram that corresponded to the mass on the current mammogram was correctly identified. The region centroid identified by the registration technique had an average distance of 2.8 ± 1.9 mm from the centroid of the radiologist-identified region. These results indicate that our new registration technique may be useful for establishing correspondence between structures on current and previous mammograms. Once such a correspondence is established an interval change analysis could be performed to aid in both detection as well as classification of abnormal breast densities. [S0094-2405(99)00612-4]

Key words: image registration, computer-aided diagnosis, computer vision, interval change, breast cancer

I. INTRODUCTION

Mammography is currently the most effective method for early breast cancer detection.[1,2] A variety of computer-aided diagnosis (CAD) techniques have recently been developed to detect mammographic abnormalities and to distinguish between malignant and benign lesions.[3–8] Knowledge from diverse areas such as signal and image processing, pattern recognition, computer vision, artificial intelligence, and neural networks has been used to develop algorithms to be implemented within a CAD scheme. Varying degrees of success for these approaches have been reported in the literature. One common feature of most of these CAD techniques is that they use a single mammogram for analysis. However, some malignancies may only manifest as a new density on mammograms without associated calcifications or masses, others distinguish themselves from benign lesions only by their relatively rapid changes in sizes. Therefore, radiologists routinely use several mammographic views along with mammograms obtained in previous years for detecting and evaluating breast lesions and for identifying interval changes. The importance of interval change analysis in mammographic interpretation has been established in clinical practice.[9,10] It can be expected that analysis of changes in mammographic features between current and previous mammograms of the patient will also be an important component of a CAD system for both the detection and the classification tasks. The ability for automated analysis of interval changes would further the ability of CAD to offer an objective second opinion. This improvement, in turn, could increase the positive predictive value of mammography, reduce the number of benign biopsies, and hence reduce both cost and patient morbidity.

While a number of CAD schemes use only a single mammogram, the simultaneous use of more than one mammogram has been under investigation for some time. Several researchers have used views of the contra-lateral breast for detecting masses and developing densities. For instance, Yin

et al.[11,12] have utilized architectural asymmetry between the right and left breasts to detect masses. While it is widely accepted that interval changes in mammographic features are very useful for both detection and classification of breast abnormalities, the development of CAD techniques to use this information has achieved limited success.[13–18] Sallam and Bowyer[13] have proposed a warping technique for mammogram registration. They manually obtained control points and calculated a mapping function for mapping each point on the current mammogram to a point on the previous mammogram. The mapping function was obtained based on local affine transformations, as well as interpolation and surface fitting techniques. A drawback of this technique is the need for manual demarcation of control points. Brzakovic *et al.*[14] have investigated a three-step method for comparison of most recent and previous mammograms. They first registered two mammograms using the method of principal axis, and partitioned the current mammogram using a hierarchical region-growing technique. The breast regions in the two mammograms were aligned with respect to each other by means of translation, rotation, and scaling. Although the technique was evaluated on a total of 64 images obtained from eight cases, this work mainly aimed toward detecting cancerous changes in breast tissue and, therefore, no quantitative analysis of registration accuracy was presented. Vujovic and co-workers[15,16] have proposed a multiple-control-point technique for mammogram registration. They first determined several control points independently on the current and previous mammograms based on the intersection points of prominent anatomical structures in the breast. A correspondence between these control points was established based on a search in a local neighborhood around the control point of interest. In a more recent publication,[17] they have evaluated their approach for establishing the correspondence between control points extracted from two mammograms using 29 temporal image pairs, and presented a qualitative evaluation based on an observer study. They have demonstrated that 91% of 103 computer-matched control points were in agreement with those matched by a radiologist. An important assumption of their work was that the distances between the control points did not change significantly between the two mammograms. However, this assumption is not necessarily a valid one. Variations in compression could potentially cause a large variation in the relative distances between the control points. Furthermore, the control points representing the intersections of elongated structures do not always have correspondences on the two mammograms. Most of these points are two-dimensional projection image of structures at different depths of an elastic and compressible three-dimensional breast. The projected intersection points can thus vary from image to image and are not invariant lankmarks. As noted by the authors, the potential control points are not points that are naturally selected by a radiologist when examining mammograms. Hence, the significance of these points is debatable.

An important factor that may limit the success of the above-mentioned techniques is that the extraction of any meaningful information from previous mammograms first requires a common frame of reference between the current and previous mammograms. Several complicating factors confound obtaining such a frame of reference. These factors include differences in breast compression and positioning between the current and previous mammograms, differences in the imaging technique between the two examinations, and changes in breast structure, size, and tissue density between the two images with patient age. As a result, the mammographic appearance of breast tissue on the current and previous mammograms of the same patient may vary considerably. Although these variabilities have not been quantified experimentally, they can be observed easily from most mammograms. Conventional registration techniques work well for applications involving rigid objects. Because of the elasticity of the breast tissue, the absence of obvious landmarks, and the large variability in the relative positions of the breast tissues projected onto the mammogram from one examination to the other, these techniques may not be optimal for registration of breast images.

In mammographic interpretation, a radiologist routinely compares the current mammogram with previous mammograms (if available) of the same view in order to detect changes in mammographic features. For example, if a mass is detected in the current mammogram, the radiologist searches for that mass in the previous mammogram to determine if this is a new or developing density. If the corresponding mass is found on the previous mammogram, then the radiologist compares the current and previous mass size and estimates if the mass has increased in size. To facilitate these comparisons, we plan to develop automated methods to detect the interval changes as a part of a computer-aided diagnostic system. As a first step, we have developed a novel method for automatic registration of lesions on temporal pairs of mammograms. In our approach, the computer emulates the search method used by many radiologists for finding corresponding structures on mammograms. The method aims at registering a small region containing a suspected mass on the most recent mammogram of the patient with one on a mammogram obtained from a previous year. Our regional registration technique involves three steps: (1) identification of a suspicious structure on the most recent mammogram, (2) initial estimation of the location on a previous mammogram of the region corresponding to the suspicious structure and the definition of a search region which encloses the object of interest on the previous mammogram, and (3) accurate identification of the location of the matched object within the search region. After the two matched lesions are identified, their characteristic features can be automatically extracted and interval changes estimated. In the present study, we focused on the development and the evaluation of the regional registration technique, rather than to solve the entire interval change analysis problem. The subsequent steps in the interval change analysis are beyond the scope of this study.

In the following sections we will provide a detailed description of our regional registration technique for temporal registration of mammograms and the results of a quantitative evaluation using a data set of 74 temporal image pairs. Although we evaluated a semiautomated version of the tech-

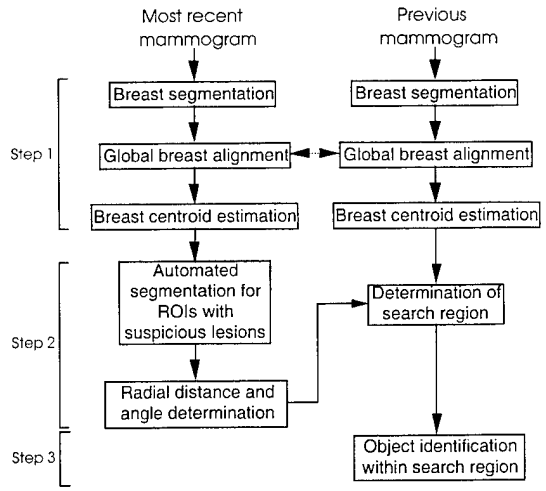


FIG. 1. Regional registration technique for determining an object on the previous mammogram which corresponds to a suspicious object on the most recent or current mammogram.

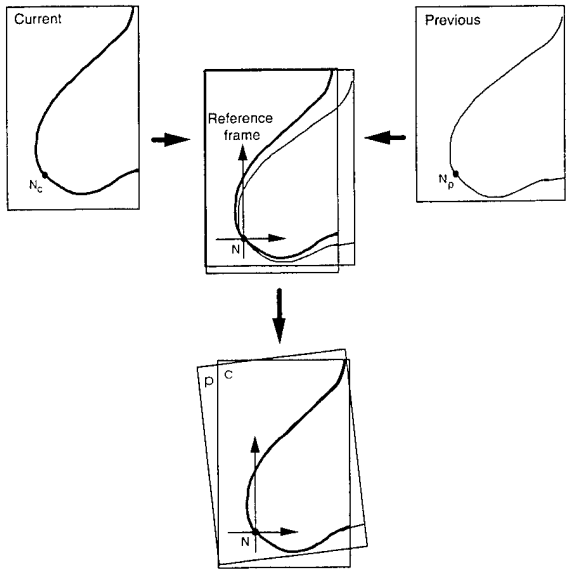


FIG. 2. Global breast alignment based on the mutual information between the two breast regions. N_c—nipple location in current mammogram, N_p—nipple location in previous mammogram, N—nipple location for both current and previous mammograms after translating them to the common frame of reference. The previous mammogram is rotated until the mutual information between the two mammograms is maximized.

nique in this preliminary study, it can be fully automated by incorporating a nipple detection step so that no user interaction will be required.

II. MATERIALS AND METHODS

A. Regional registration and mammogram correspondence

As the term indicates, regional registration is a local rather than a global registration technique. It is a multistep procedure and utilizes computer-detected objects in the most recent (hereafter termed current) mammogram. In the context of this paper, a current mammogram is either the latest mammogram of the patient, or the latest mammogram before biopsy. The detected objects could be either true masses (benign or malignant) or false positives (normal breast structures). Regional registration then finds a matching object on a previous mammogram. The three major steps in regional registration are illustrated in Fig. 1 and details of the technique are described below.

In the first step of regional registration, the breast region is segmented from the background on both the current and the previous mammograms. For this purpose we have used a breast boundary detection algorithm previously developed in our laboratory.[19,20] This algorithm could successfully track the breast boundaries in over 90% of the 1000 mammograms in a previous study. It performed reliably on all the images in our database. After extracting the breast border from the mammogram, the location of the nipple is estimated on both the current and the previous mammograms. Any automated method[21,22] can be used for finding the nipple location. However, in this study, the nipple location was manually identified by a radiologist for all images in our data set. The breast border and the nipple location now form the basis of a global breast alignment (GBA) procedure illustrated in Fig. 2. Since the sizes and the orientations of the two images could vary between the current and previous mammograms, a common frame of reference is needed. The GBA procedure has been devised specifically to provide such a frame of reference. We first define a new frame of reference with the nipple location on the current mammogram (N_c) as the origin. The previous mammogram is translated so that its nipple location (N_p) aligns with the origin in the common frame of reference as shown in Fig. 2. Using the origin as the pivot point, we rotate the previous mammogram to align the breast regions in the two images.

We have evaluated two different methods for estimation of the optimum rotation angle. The first method is based on maximization of the overlap area, and the second method is based on maximization of the mutual information (MI)[23,24] between the two segmented breast regions. To determine the MI, we first rescale the breast portion of both mammograms to a 0–255 gray scale. For a given rotation angle θ, the two-dimensional (2D) histogram $h_\theta(i,j)$ of the gray levels for the corresponding pixels on the current mammogram and the previous mammogram is constructed. Here i refers to the gray level on the current mammogram and j refers to the gray level on the previous mammogram rotated by an angle θ. The probability density of the gray scale co-occurrences is estimated from the 2D histogram as

$$f_\theta(i,j)=\frac{h_\theta(i,j)}{\Sigma_{m,n}h_\theta(m,n)}, \tag{1}$$

where $0\leqslant i,j\leqslant 255$, $0\leqslant m,n\leqslant 255$. The mutual information (MI_θ) between the two images for a specific rotation angle θ is computed as

$$\mathrm{MI}_\theta=\sum_{i,j} f_\theta(i,j)^* \log_2\left\{\frac{f_\theta(i,j)}{\Sigma_m f_\theta(i,m)\Sigma_n f_\theta(n,j)}\right\}. \tag{2}$$

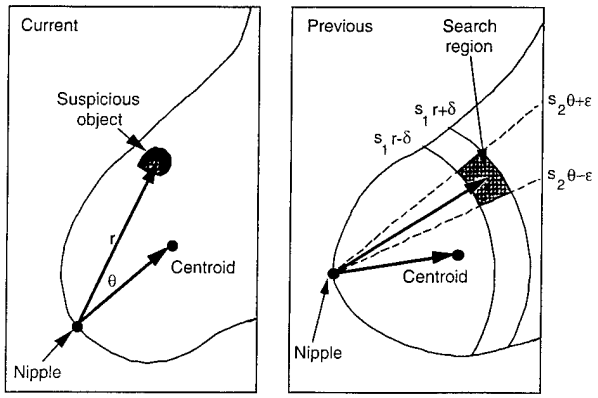


FIG. 3. Polar coordinate system defined using the nipple location and the nipple–centroid axis. The search region for finding a matching object on the previous mammogram is shown as the shaded region.

The above-mentioned procedure is repeated for several rotation angles and the angle θ_{max} which provides the maximum mutual information is chosen for global breast alignment of the previous mammogram and the current mammogram. Note that while the area overlap method for GBA uses the binary image after segmentation, the MI-based method uses the original gray scale image. The effects of the two methods on the accuracy of regional registration will be discussed later in Sec. IV. Once the two images are aligned in the common frame of reference, the centroid of the breast region is estimated, and the nipple–centroid axis is defined for both mammograms. For comparison we also show in Sec. III regional registration results based on computing the centroids of the two breast regions without global breast alignment. The nipple–centroid axis forms the basis for the second step of regional registration.

In the second step, suspicious regions are automatically segmented from the breast region on the current mammogram. This can be accomplished by using a density-weighted contrast enhancement (DWCE) technique[25] previously developed in our laboratory. While the use of the DWCE technique is not critical for regional registration, it does help automate the entire procedure. Alternatively, a radiologist can manually identify a suspicious object or a region of interest on the current mammogram and the regional registration technique can be used to identify a corresponding region on the previous mammogram. Once suspicious objects have been identified on the current mammogram, the centroid of each object is estimated. A polar coordinate system is then defined using the nipple as the origin and the nipple–centroid axis as the 0° axis on both images. This is illustrated in Fig. 3. The location of the centroid of a suspicious object on the current mammogram is determined as (r,θ). We then compute two scale factors—the radial scale factor s_1 and the angular scale factor s_2. These scale factors have been devised to provide a first-order correction for factors such as breast compression differences between the current and previous mammograms, differences in image magnification and size, and changes in overall breast shape between the two images. The radial scale factor s_1 is estimated as the ratio of the nipple–centroid distances on the previous and current images. The angular scale factor s_2 is estimated as the ratio of the angular width of the breast on the previous image at radius $s_1 r$ to that on the current image at radius r. The initial estimate of the corresponding location of the suspicious object on the previous mammogram is then obtained as $(s_1 r, s_2 \theta)$.

Using the initial estimate of the centroid of the object on the previous mammogram, we can define a fan-shaped search region bounded by $s_1 r \pm \delta$ and $s_2 \theta \pm \epsilon$ as illustrated in Fig. 3. The object found on the current mammogram is then used as a template to search for a matching object in the search region on the previous mammogram. The size of the search region (defined by δ and ϵ) depends on the variability between mammograms obtained from one examination to the other. Since it is difficult to predict the variability of an elastic and deformable object such as the breast by any analytical method, we have determined this variability experimentally from the mammograms in our data set. The variation in compression can cause a change in the relative locations of various breast structures on these images as well as a rotation of the breast boundary with respect to the fixed image coordinates. By relating the position of a breast structure to the corresponding nipple–centroid axis, and by performing a search in the corresponding search region, we can reduce the effect of this variability. In this study we have estimated the size of the search region required to enclose all corresponding objects on the previous mammogram using ground truth objects identified on the previous mammograms by a radiologist. The distance of the initial estimate of the center of the search region from the centroid of the ground truth object was also estimated.

The third and final step in the regional registration procedure involves a systematic search to identify a corresponding structure within the fan-shaped search region on the previous mammogram. In this study we have evaluated two different search criteria. The first criterion is based on gray scale template matching. A rectangular gray scale template centered on the mass centroid is extracted from the current mammogram. The choice of the size of the template region can affect the accuracy of the registration technique. The minimum required size of a rectangular template is, of course, a rectangular region which encloses the mass exactly. However, one can also include a small portion of the background region in the template. We have analyzed the performance of our algorithm using two different sizes for this template. The first includes a 1-pixel-wide background region all around the boundary of the suspicious object while the second includes a 5-pixel-wide background region. For each pixel (i,j) in the fan-shaped region on the previous mammogram, a region of interest (ROI) centered on the pixel and of the same size as the mass template is extracted. We denote the (m,n)th pixel in the gray scale template extracted from the current mammogram as $p(m,n)$ and that from the ROI obtained from the fan-shaped region as $q_{i,j}(m,n)$. A correlation measure defined as

$$C_{i,j}=\frac{\Sigma_{m,n}(p(m,n)-\bar{p})(q_{i,j}(m,n)-\bar{q})}{\sqrt{(\Sigma_{m,n}(p(m,n)-\bar{p})^2)(\Sigma_{m,n}(q_{i,j}(m,n)-\bar{q})^2)}} \quad (3)$$

is then obtained for each pixel (i,j) within the search region on the previous mammogram. Here the summation is performed over the mass template, and $\bar{p}$ and $\bar{q}$ denote the average pixel values in the template and ROI, respectively. The correlation values in the search region are then smoothed by a 3×3 averaging kernel to reduce fluctuations. The final estimate of the location of the mass centroid on the previous mammogram is obtained as the location corresponding to maximum correlation. The second search criterion is based on maximizing the mutual information between the mass template and the ROI extracted from within the search region. The MI approach is similar to that described earlier for alignment of the breast regions, except that the regions to be matched are limited to the size of the mass template.

Once a corresponding structure is found on the previous mammogram for a suspicious object on the current mammogram, it can be used for an interval change analysis within a CAD scheme, as we have shown in an independent study.[26] If the search procedure in the fan-shaped region does not yield a corresponding region, then the suspicious object on the current mammogram can be considered as a newly developed density. Objects for which no corresponding object can be found on the previous mammogram can be analyzed with methods designed for single images in an overall CAD scheme. Note that in this study the search techniques are structured in a way to always determine a matching object. Search criteria to identify new densities will be developed in future studies.

B. Image acquisition and data set

The data set for this study consisted of 127 images obtained from the files of 34 patients who had undergone biopsy at the University of Michigan. From these 127 mammograms, 74 temporal pairs of images were obtained. The current mammogram of each temporal pair exhibited a biopsy-proven mass. All previous mammograms in the 74 temporal pairs contained a mass, a structure, or a density which the radiologist could match to the mass detected in the corresponding current image. Since some patient files contained a sequence of mammograms over three years, the number of temporal pairs was larger than half the number of images. The 74 temporal image pairs were comprised of 43 cranio-caudal views and 31 mediolateral-oblique views.

The mammograms of 20 temporal pairs were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 0.1 mm$\times$0.1 mm and with 12 bit resolution. The digitizer was calibrated so that the gray values were linearly and inversely proportional to the optical density (OD) within the range of 0.1–2.8 OD units, with a slope of -0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of this digitizer was 0–3.5. The mammograms of the remaining 54 temporal pairs were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 0.05 mm$\times$0.05 mm and with 12 bit resolution. This digitizer was calibrated so that the gray values were linearly and inversely proportional to the OD within the range 0–4 OD units, with a slope of -0.001 OD/pixel value. All images were subsequently reduced to 0.8 mm resolution by averaging adjacent 8×8 pixels (20 pairs) or 16×16 pixels (54 pairs). Since the same digitizer was used for digitizing all films of the same case, the differences in the digitizers would have no effect on the analysis of each image pair. Given the small differences between the two laser digitizers and the large differences in the imaging technique and in the breast appearance from one case to another, it could be expected that the use of cases collected with the two different digitizers would not affect the evaluation of the registration technique.

While the regional registration technique can be used for determining a corresponding structure or region for any structure (both false positives and masses) in the breast, in this study we have analyzed its accuracy on biopsy-proven masses alone. The location of the mass on the current mammogram was identified by an MQSA-certified radiologist experienced in breast imaging. The radiologist manually identified the corresponding region on the previous mammogram and the nipple location on both the current and the previous mammograms using an interactive image analysis tool on a UNIX workstation. For each current mammogram, the boundary of the mass was manually delineated by the radiologist using an image display program developed in our laboratory. A bounding box enclosing the corresponding object on the previous mammogram was provided by the radiologist for each of the masses. Each mass as well as the corresponding structure on the previous mammogram was rated for its visibility on a scale of 1–10, where the rating of

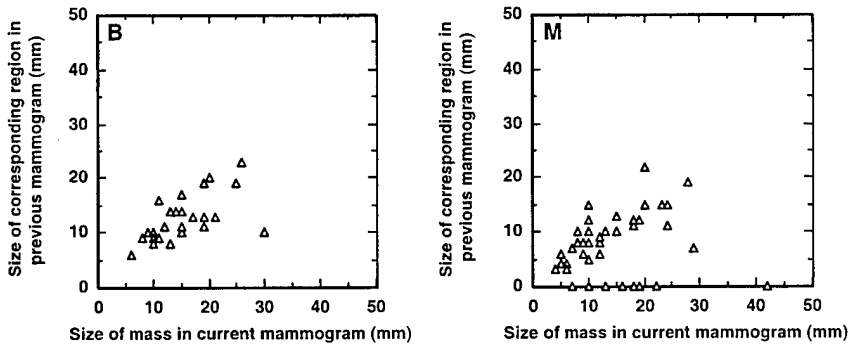


FIG. 4. Distribution of the size of the mass on the current mammogram with respect to the size of the corresponding structure on the previous mammogram as estimated by an experienced breast radiologist for benign (B) and malignant (M) cases in the data set.

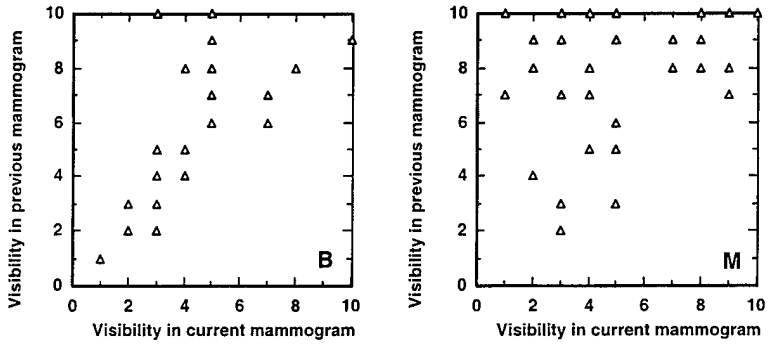


FIG. 5. Distribution of the visibility of the mass on the current mammogram with respect to the visibility of a corresponding structure on the previous mammogram as rated by an experienced breast radiologist for benign (B) and malignant (M) cases. In this rating scale the visibility of the masses decreases from 1 to 10 with 10 being the least visible. The total number of points in these two graphs is less than the total number of mammogram pairs in our database, because mammogram pairs with the same rating appear as a single point.

1 corresponded to the most visible category. The size of the mass on the current mammogram as well as the size of the corresponding structure on the previous mammogram was also provided by the radiologist. For previous mammograms on which the radiologist could not identify a distinct mass, the "mass" size was given a size of 0 mm. The parenchymal density was rated based on the BIRADS lexicon. The distributions of the size and visibility ratings for benign and malignant cases in this data set are shown in Figs. 4 and 5.

C. Evaluation of registration accuracy

The bounding box enclosing the corresponding object on the previous mammogram provided by the radiologist was used as the "ground truth" to evaluate the accuracy of the regional registration technique. We have used two different measures for assessing registration accuracy. The first measure quantifies whether the corresponding region is correctly identified by the registration algorithm. This measure is computed simply as the number of cases in which the estimated centroid location of the mass on the previous mammogram is inside the bounding box provided by the radiologist. The second measure quantifies the error in the estimate of the corresponding region on the previous mammogram and is defined as the Euclidean distance between the estimated centroid of the corresponding region and the center of the bounding box provided by the radiologist. Together these two measures answer the questions: (a) does regional registration work? (b) how well does the technique perform in matching structures between the current and previous mammograms? In Sec. III we provide the results of regional registration with and without global breast alignment and using both correlation and mutual information as the search criterion in step 3.

III. RESULTS

To provide the reader with a qualitative idea of algorithm performance we first illustrate the intermediate results at various stages of the algorithm. Then the results of each of the three steps of the algorithm are presented with an analysis of the dependence of the performance on various algorithm parameters. Also presented is an analysis of the accuracy of regional registration using the error measures defined in Sec. II C. In the following sections, the term "initial estimate" refers to the estimate of the center of the search region in step 2 of regional registration. The term "final estimate" refers to the outcome of the search procedure adopted in step 3 and represents the overall result of regional registration.

A. Intermediate results of regional registration

Figures 6–8 show an example of the intermediate and final results of applying the regional registration technique to a temporal pair of mammograms. The original digitized mammograms—current and previous—with the automati-

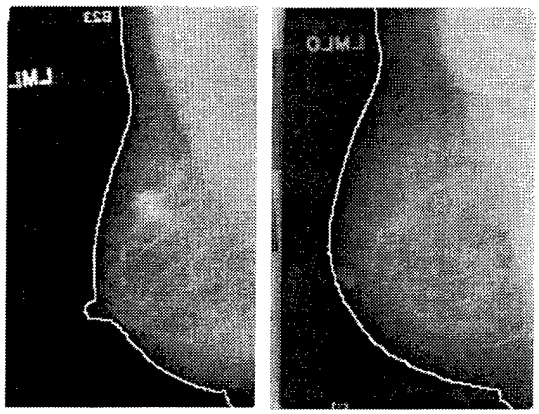


FIG. 6. Left—most recent or current mammogram. Right—previous mammogram. The breast images are superimposed with the breast borders detected by a breast boundary tracking algorithm.

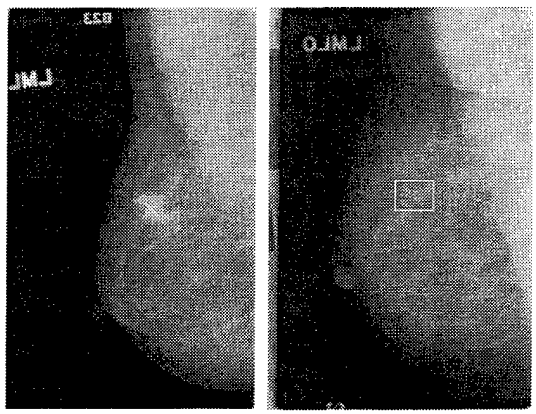


FIG. 7. Left—location of the mass on the current mammogram. Right—radiologist-identified region on previous mammogram corresponding to the mass on the current mammogram.

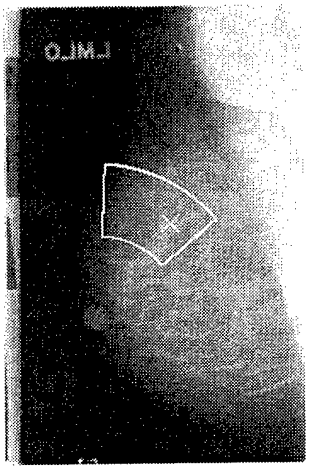


FIG. 8. The fan-shaped search region on the previous mammogram. The initial computer estimate of the centroid location of the region corresponding to the mass is at the center of the search region. The final estimate of the centroid of the corresponding region (indicated by X) is obtained by using the correlation criterion within the fan-shaped search region.

cally tracked breast boundaries superimposed, are shown in Fig. 6. The location of the mass on the current mammogram is shown in Fig. 7 along with the corresponding radiologist-identified region on the previous mammogram. Figure 8 shows the fan-shaped search region on the previous mammogram estimated in step 2 of regional registration. The initial estimate is at the center of this search region which is to be used in step 3 for localization of the corresponding mass. The centroid location of the corresponding object estimated by the algorithm using the correlation measure as the search criterion is also shown in Fig. 8.

B. Initial estimates and search regions

Figure 9 shows histograms of the Euclidean distance between the initial estimate of the centroid location of the corresponding structure on the previous mammogram and the center of the bounding box provided by the radiologist. For the 74 temporal image pairs used in this data set, the average Euclidean distance error of the initial estimate was 10.5 mm (std. dev. 6.4 mm) without the GBA procedure and 9.8 mm (std. dev. 6.0 mm) with the GBA procedure. The overall accuracy was 46% in both cases, i.e., in 34 of the 74 temporal image pairs the initial estimate was inside the ground-truth bounding box. Based on observation of the radial deviation errors and the angular deviation errors (defined in Sec. IV) in Figs. 10 and 11, a search region defined by $\epsilon = 0.35 + 5/r$ rad and $\delta = 20$ mm with GBA ($\delta = 25$ mm for no GBA), where r is the radial distance from the nipple, was used for the evaluation of the local search criteria used in step 3 of regional registration.

C. Local search criteria and final estimates

Figure 12 shows the histograms of the Euclidean distance errors of the final estimate of the corresponding structure using the correlation measure as the search criterion. Table I summarizes the results along with the average Euclidean distance errors and standard deviations using both the correlation and the mutual information search criteria and with and without the GBA procedure. The average Euclidean distance errors and deviations for the cases where the final estimate is inside the ground-truth region identified by the radiologist and the cases where it is outside are also listed separately. Regional registration incorporating the GBA procedure and using correlation as a search criterion has an accuracy of 85%. In 63 of the 74 temporal image pairs, the final estimate of the location of the corresponding region was inside the radiologist-identified ground-truth region. The use of mutual information as a search criterion yielded an accuracy of 74% (55 out of 74 temporal pairs). The average Euclidean distance error for regional registration incorporating GBA and correlation was 4.7 mm (std. dev. 5.8 mm) for all 74 temporal pairs and 2.8 mm (std. dev. 1.9 mm) in 85% (63/74) of the temporal pairs. Use of mutual information as a search criterion in step 3 results in values of 7.2 mm (std dev. 8.6 mm) and 3.0 mm (std. dev. 2.0 mm), respectively, for the same quantities.

IV. DISCUSSION

A. Initial estimates and search regions

From the histograms of Fig. 9, we observe that the use of the GBA procedure results only in a marginal improvement in the initial estimate, if the Euclidean distance error is the only measure considered. However, the GBA procedure has a significant effect in reducing the size of the search region required for regional registration. In order to compute the required sizes (δ and ϵ in Fig. 3) of the search region, we computed two quantities—the radial distance deviation and the angular deviation—using the initial estimate obtained from step 2 for the 74 temporal image pairs. The radial distance deviation is defined as the absolute difference between $s_1 r$ and r_c, where r_c is the radial distance of the center of the ground-truth region from the nipple location on the pre-

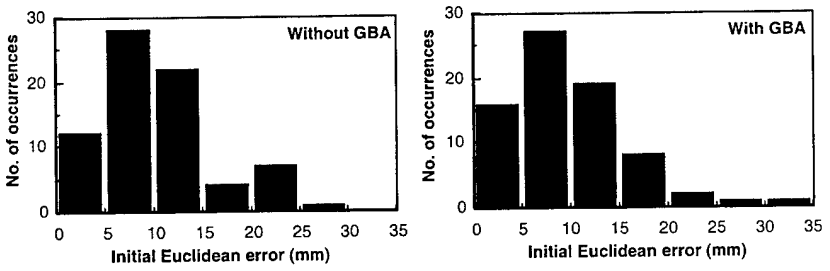


FIG. 9. Histograms of Euclidean distance between the initial estimate of the centroid location of the corresponding object and the center of the radiologist-identified object on the previous mammogram with and without GBA.

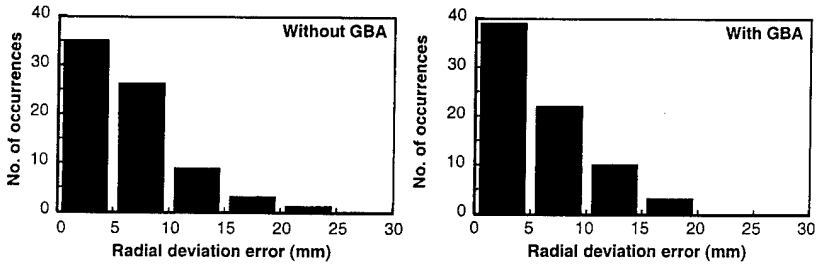


FIG. 10. Histograms of radial distance deviation between the initial estimate of the centroid location of the corresponding object and the center of the radiologist-identified object on the previous mammogram with and without GBA.

vious mammogram. The histograms of radial distance deviations for the 74 temporal image pairs with and without the GBA procedure are shown in Fig. 10. An important observation is that a δ value of 25 mm is needed to include the centers of the ground-truth structures if the GBA procedure is not used in step 1. The use of the GBA procedure results in a decrease in the value of δ to 20 mm. This decrease helps significantly increase the overall accuracy of the regional registration as discussed below.

In Fig. 11 the angular deviation of the initial estimate is plotted against the radial distance of the centers of the ground-truth regions on the previous mammogram. The angular deviation ϵ is defined as $s_2\theta - \theta_c$ where θ_c is the angle between the nipple-ground-truth center vector and the nipple-centroid axis. In an earlier study[27] using both false positives and masses, we have observed that the value of ϵ needed to include the center of the ground-truth region decreases with distance from the nipple, i.e., increases with distance from the chest wall. This may be attributed to the increased deformability of the breast tissue closer to the nipple compared to the tissue closer to the chest wall. This indicates that a possible approach to take into account this variability is to incorporate a variable ϵ, one which is inversely proportional to the radial distance r from the nipple. For the data set in this study, we have investigated several forms for this dependence all of which fit under the general model

$$\epsilon = \epsilon_{th} + K/r.$$

Here ϵ_{th} and K are two constants which affect the form of the dependency. Based on our observation of the angular deviations for the entire data set of 74 temporal pairs we have chosen $\epsilon_{th} = 0.35$ rad and $K = 5$ rad-mm. As can be seen from Fig. 11, with these values of ϵ_{th} and K, all of the centers of the ground-truth regions are within the search region. Therefore, a search region defined by $\epsilon = 0.35 + 5/r$ rad, and $\delta = 20$ mm (if GBA was applied) or $\delta = 25$ mm (if GBA was not applied) was used for evaluation of the local search criteria used in step 3 of regional registration.

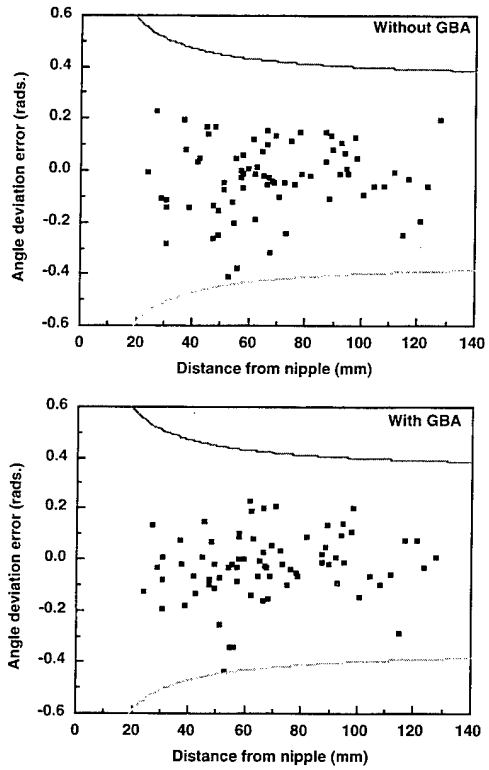


FIG. 11. Angular deviation between the initial estimate of the centroid location of the corresponding object and the center of the radiologist-identified object on the previous mammogram with and without GBA. Also shown are the bounding lines defined using $\epsilon = 0.35 + 5/r$ rad.

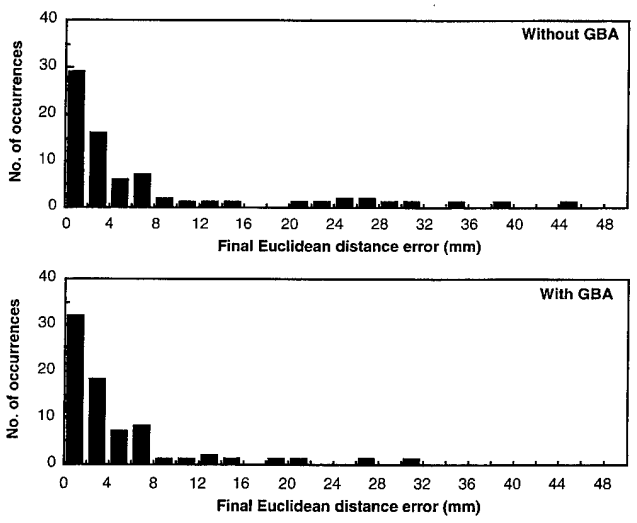


FIG. 12. Histograms of Euclidean distance error for corresponding regions estimated by regional registration using the correlation measure in step 3 with and without GBA. This error is defined as the Euclidean distance between the centroid location of the estimated corresponding region and the center of the radiologist-identified ground-truth corresponding region on the previous mammogram.

TABLE I. Accuracy of regional registration using correlation measure and mutual information measure in step 3 with and without global breast alignment (GBA) and using a 1-pixel-wide background region for the template from the current mammogram. Correct estimates are the cases where the estimated centroid location was within the bounding box of the radiologist-identified object location.

Method	Accuracy	Overall average error (mm)	Average error (mm) for correct estimates	Average error (mm) for incorrect estimates
Correlation without GBA	77% (57/74)	7.4±10.2	2.8±2.0	22.9±11.5
Mutual information without GBA	68% (50/74)	8.8±10.5	3.0±2.0	20.7±11.1
Correlation with GBA	85% (63/74)	4.7±5.8	2.8±1.9	15.7±8.3
Mutual information with GBA	74% (55/74)	7.2±8.6	3.0±2.0	19.4±8.9

B. Local search criteria and final estimates

We have evaluated the use of correlation and mutual information as the local search criteria. From Table I we observe that the GBA procedure results in a higher accuracy irrespective of the search criterion. While the use of mutual information as a search criterion performs reasonably well by itself (74% accuracy with an average error of 7.2 mm) the use of correlation measure was observed to result in more accurate registration. For the images in this data set, the correlation measure outperformed the mutual information measure irrespective of whether the breast centroids were computed with or without the GBA procedure.

A few observations on the 11 cases where the final estimate was outside the radiologist-identified ground-truth corresponding region are in order. In 7 of the 11 cases although the radiologist did provide a region corresponding to the mass on the current mammogram, the corresponding structure on the previous mammogram was very subtle (visibility rating 8 or higher) with indistinct boundaries. The radiologist could only estimate the region where the mass would develop rather than the mass itself, so the truth was uncertain. In one of the remaining 4 cases, the mass was an architectural distortion in the current mammogram. In a second (benign) case the mass shape had changed considerably. Upon consultation of the pathology report, the radiologist concluded that the mass was a benign cyst which had been aspirated in the previous year resulting in a substantial change in its shape. In the third case, the proximity of the mass to the chest wall resulted in it being incompletely imaged in the previous year compared to the current year. In such cases the correlation measure of a neighboring breast structure would tend to be higher than that of the corresponding structure. In the fourth case, an overlap of two vessels was identified as corresponding to the mass on the current mammogram while the region corresponding to the mass was observed to be extremely subtle. In almost all of the 11 cases the proximity of the corresponding region to a dense structure combined with the subtle nature of the structure on the previous mammogram render the correlation measure ineffective in establishing correspondence. However, in clinical practice, these masses will likely be categorized as a newly developed density. Criteria to distinguish a newly developed density will be investigated in further studies.

C. GBA: Area overlap vs mutual information

For the images used in this study, the result of the GBA procedure based on maximizing the area overlap between the breast regions in the two images of a temporal pair is comparable to that based on maximizing the mutual information. However, our observation is that the mutual information criterion is preferable to the area overlap criterion. The area overlap measure suffers from the drawback that if the breast region in one of the mammograms is uniformly smaller than that in the other, i.e., the breast edge in one is completely within the breast edge in the other, then there is no unique rotation angle at which the area overlap is maximized. Although the range of rotation angles over which local maxima of the area overlap occur is small, the resulting estimate of the rotation angle for GBA may be suboptimal. The use of mutual information, however, results in a single unique rotation angle at which MI is maximized. In any case, as discussed earlier, the use of the GBA procedure before computing the breast centroid results in a reduction in the size of the search region. A smaller search region reduces the likelihood that the mass template is matched to an incorrect structure and, therefore, increases the accuracy and reduces the Euclidean distance error.

D. Template size, scale factors, and computation times

The size of the background region in the gray scale template extracted from the current mammogram affects registration accuracy. For the 74 temporal pairs in this data set, the best performance was observed when a 1-pixel-wide background region was included all around the boundary of the mass template. A 5-pixel-wide background region resulted in a decrease in accuracy and an increase in the average Euclidean distance error. The accuracy progressively decreased and the Euclidean distance error increased with an increase in the size of the background region in the template. Figure 13 shows the distributions of the radial and angular scale factors for the images used in this study. The radial scale factor s_1 ranged from 0.94 to 1.05 for this data set. Use of s_1 reduced the size of the search area by decreasing the required value for δ. The angular scale factor s_2 was very close to 1 in all cases and did not seem to make any major difference for the images in this data set. On a final note the computation time required for regional registration incorporating correlation was on the average 2 s without GBA and 4 s with GBA on a UNIX workstation (DEC AlphaStation 600 series).

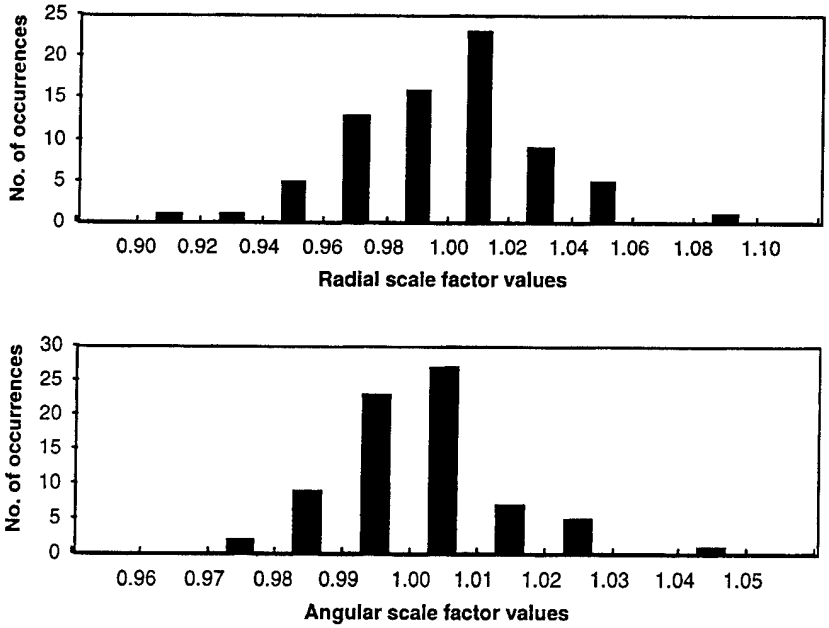


FIG. 13. Histograms of the radial scale factor and the angular scale factor for 74 temporal image pairs. The radial scale factor s_1 is estimated as the ratio of the nipple–centroid distances on the previous and current images. The angular scale factor s_2 is estimated as the ratio of the angular width of the breast on the previous image at radius $s_1 r$ to that on the current image at radius r.

V. CONCLUSIONS

Radiologists are interested in determining any local changes in breast tissue over time which may indicate a developing cancer. We have developed a novel regional registration technique for temporal registration of mammograms. This technique could become an important component of a CAD scheme for mammographic analysis. Unlike other techniques found in the literature, our regional registration technique does not depend on the identification of landmark structures or control points on the mammograms. It is based on a search technique that many radiologists use and has proven to be successful in mammographic interpretation. After corresponding objects are found, they can be analyzed for interval changes in a CAD scheme. Our preliminary results indicate that the regional registration technique is promising in identifying corresponding regions from temporal mammographic pairs. In 85% (63/74) of the cases the regional registration technique correctly identified the corresponding region in the previous mammogram. For these 63 cases, it is highly encouraging to note that the estimated location of the region corresponding to the mass in the current mammogram was less than 3 mm on the average from radiologist-identified corresponding locations.

Areas for future work include the development of an automated technique for identifying the nipple location on the mammograms, investigation of other local search criteria such as Fourier descriptors and shape-invariant moments to be used in the fan-shaped search region, adaptive methods for determining the size of the search region, criteria for identifying newly developed densities, application of regional registration to false positives as well as masses, and studies with a large data set to investigate the robustness of the regional registration technique. It may be noted that the regional registration technique may also be applicable to other related registration problems, such as the registration of left and right mammograms.

ACKNOWLEDGMENTS

This work is supported by a Career Development Award from the USAMRMC Grant No. DAMD 17-98-1-8211, USPHS Grant No. CA 48129, and USAMRMC Grant No. DAMD 17-96-1-6254. The content of this publication does not necessarily reflect the position of the government, and no official endorsement of any equipment or product of any companies mentioned in the publication should be inferred.

[a]Electronic mail: chanhp@umich.edu

[1] S. A. Feig and R. E. Hendrick, "Risk, Benefit, and Controversies in Mammographic Screening," in *Syllabus: A Categorical Course in Physics Technical Aspects of Breast Imaging*, edited by A. G. Haus and M. J. Yaffe (Radiological Society of North America, Oak Brook, IL, 1993).

[2] C. Byrne, C. R. Smart, C. Cherk, and W. H. Hartmann, "Survival advantage differences by age: Evaluation of the extended follow-up of the Breast Cancer Detection Demonstration Project," Cancer (N.Y.) **74**, 301–310 (1994).

[3] Y. Wu, K. Doi, M. L. Geiger, and R. M. Nishikawa, "Computerized detection of clustered microcalcifications in digital mammograms: Applications of artificial neural networks," Med. Phys. **19**, 555–560 (1992).

[4] H. P. Chan *et al.*, "Improvement in radiologists' detection of clustered microcalcifications: The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).

[5] S. M. Lai, X. Li, and W. F. Bischof, "On techniques for detecting circumscribed masses in mammograms," IEEE Trans. Med. Imaging **8**, 377–386 (1989).

[6] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," IEEE Trans. Med. Imaging **12**, 664–669 (1993).

[7] W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).

[8] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminant analysis in texture feature space," Phys. Med. Biol. **40**, 857–876 (1995).

[9] L. W. Bassett, B. Shayestehfar, and I. Hirbawi, "Obtaining previous mammograms for comparison: Usefulness and costs," Am. J. Roentgenol. **163**, 1083–1086 (1994).

[10] E. A. Sickles, "Periodic mammographic follow-up of probably benign lesions: Results in 3183 consecutive cases," Radiology **179**, 463–468 (1991).

[11] F. F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," Med. Phys. **18**, 955–963 (1991).

[12] F. F. Yin, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique," Med. Phys. **21**, 445–452 (1994).

[13] M. Sallam and K. Bowyer, "Detecting abnormal densities in mammograms by comparison with previous screenings," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).

[14] D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic, and K. Fogarty, "Mammogram analysis by comparison with previous screenings" in Ref. 13.

[15] N. Vujovic, D. Brzakovic, and K. Fogarty, "Detection of cancerous changes in mammograms using intensity and texture measures," Proc. SPIE **2434**, 37–47 (1995).

[16] N. Vujovic, P. Bakic, and D. Brzakovic, "Detection of potentially cancerous signs by mammogram followup," in Ref. 13.

[17] N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," IEEE Trans. Image Process. **6**, 1388–1399 (1997).

[18] W. K. Zouras, M. L. Giger, P. Lu, D. E. Wolverton, C. J. Vyborny, and K. Doi, "Investigation of a temporal subtraction scheme for computerized detection of breast masses in mammograms," in Ref. 13.

[19] A. R. Morton, "Design of an x-ray beam equalization filter for mammographic imaging," M.S. thesis, Department of Environmental and Industrial Health, University of Michigan, 1996.

[20] A. R. Morton, H. P. Chan, and M. M. Goodsitt, "Automated model-guided breast segmentation algorithm," Med. Phys. **23**, 1107–1108 (1996).

[21] A. J. Mendez, P. G. Tahoces, M. J. Lado, M. Souto, J. L. Correa, and J. J. Vidal, "Automatic detection of breast border and nipple in digital mammograms," Comput. Methods Programs Biomed. **49**, 253–262 (1996).

[22] R. Chandrasekhar and Y. Attikiouzel, "A simple method for automatically locating the nipple on mammograms," IEEE Trans. Med. Imaging **16**, 483–494 (1997).

[23] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).

[24] A. Maintz, E. Meijering, and M. Viergever, "General multimodal elastic registration based on mutual information," Proc. SPIE **3338**, 144–154 (1998).

[25] N. Petrick, H. P. Chan, D. Wei, B. Sahiner, M. A. Helvie, and D. D. Adler, "Automated detection of breast masses on mammograms using adaptive contrast enhancement and tissue classification," Med. Phys. **23**, 1685–1696 (1996).

[26] S. Sanjay-Gopal, H. P. Chan, B. Sahiner, N. Petrick, T. Wilson, and M. Helvie, "Evaluation of interval change in mammographic features for computerized classification of malignant and benign masses," Radiology **205(P)**, 216 (1997).

[27] S. Sanjay-Gopal, H. P. Chan, N. Petrick, T. Wilson, B. Sahiner, M. Helvie, and M. Goodsitt, "A regional registration technique for automated analysis of interval changes of breast lesions," Proc. SPIE **3338**, 118–131 (1998).

Classifier design for computer-aided diagnosis: Effects of finite sample size on the mean performance of classical and neural network classifiers

Heang-Ping Chan[a] and Berkman Sahiner
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

Robert F. Wagner
Center for Devices and Radiology Health, FDA, Rockville, Maryland 20852

Nicholas Petrick
Department of Radiology, University of Michigan, Ann Arbor, Michigan 48109-0030

(Received 14 June 1999; accepted for publication 16 September 1999)

Classifier design is one of the key steps in the development of computer-aided diagnosis (CAD) algorithms. A classifier is designed with case samples drawn from the patient population. Generally, the sample size available for classifier design is limited, which introduces variance and bias into the performance of the trained classifier, relative to that obtained with an infinite sample size. For CAD applications, a commonly used performance index for a classifier is the area, A_z, under the receiver operating characteristic (ROC) curve. We have conducted a computer simulation study to investigate the dependence of the mean performance, in terms of A_z, on design sample size for a linear discriminant and two nonlinear classifiers, the quadratic discriminant and the backpropagation neural network (ANN). The performances of the classifiers were compared for four types of class distributions that have specific properties: multivariate normal distributions with equal covariance matrices and unequal means, unequal covariance matrices and unequal means, and unequal covariance matrices and equal means, and a feature space where the two classes were uniformly distributed in disjoint checkerboard regions. We evaluated the performances of the classifiers in feature spaces of dimensionality ranging from 3 to 15, and design sample sizes from 20 to 800 per class. The dependence of the resubstitution and hold-out performance on design (training) sample size (N_t) was investigated. For multivariate normal class distributions with equal covariance matrices, the linear discriminant is the optimal classifier. It was found that its A_z-versus-$1/N_t$ curves can be closely approximated by linear dependences over the range of sample sizes studied. In the feature spaces with unequal covariance matrices where the quadratic discriminant is optimal, the linear discriminant is inferior to the quadratic discriminant or the ANN when the design sample size is large. However, when the design sample is small, a relatively simple classifier, such as the linear discriminant or an ANN with very few hidden nodes, may be preferred because performance bias increases with the complexity of the classifier. In the regime where the classifier performance is dominated by the $1/N_t$ term, the performance in the limit of infinite sample size can be estimated as the intercept ($1/N_t = 0$) of a linear regression of A_z versus $1/N_t$. The understanding of the performance of the classifiers under the constraint of a finite design sample size is expected to facilitate the selection of a proper classifier for a given classification task and the design of an efficient resampling scheme. © *1999 American Association of Physicists in Medicine.*
[S0094-2405(99)00212-6]

Key words: computer-aided diagnosis, classifier design, linear classifier, quadratic classifier, neural network, sample size, feature space dimensionality, ROC analysis

I. INTRODUCTION

With the advent of digital imaging modalities, computer-aided diagnosis (CAD) is becoming an important area of research in medical imaging. A CAD algorithm can detect abnormalities and classify disease or normal cases based on image and/or patient information, and thus provide a second opinion to the radiologist in the detection or diagnostic decision making process.

Design of classifiers that can accurately distinguish normal and abnormal features is a critical step in the development of CAD algorithms. It has been shown that the performance of a classifier for unknown cases depends on the sample size used for training.[1] When a finite design (training) sample size is used, the performance is pessimistically biased in comparison to that obtained from an infinitely large design sample. In order to design a classifier with a performance generalizable to the population at large, one has to use a sufficient number of case samples that are representative of the population. However, the availability of case samples is often limited in medical imaging research. It is therefore important to study the sample-size dependence of different classifiers and determine the most efficient way of training a classifier, under the constraint of a finite sample size.

 0094-2405/99/26(12)/2654/15/$15.00

We note that the concept of generalizability may be used in several technical senses when assessing the performance of a classifier: one with respect to mean classifier performance, the other with respect to the variance of classifier performance. In many classifier design problems, one is most interested in investigating if the mean performance of a classifier estimated from a given set of finite design samples can be generalized to classification performance with unknown test samples drawn from the same population of cases. The generalizability in this regard can be observed from the biases of the mean performances in the finite design set and in the test set in comparison to the optimal performance estimated from an infinite design set. The bias in the mean performance of different classifiers under various input conditions is the subject of investigation in this study. We will discuss further other interpretation of generalizability in the Discussion section of this paper.

A number of investigators have studied the finite-sample-size problem[1–9] Fukunaga[1,3] derived a general formulation for the bias and variance of a function, f, which is to be estimated from the available samples. When f is a nonlinear function of the mean vectors and covariance matrices of two feature distributions, it has been shown that a bias results from the nonlinear propagation of the finite-sample variances in the estimates of the mean vectors and covariance matrices of the distributions through this function. For multivariate-normal data, these variances are proportional to $1/N_t$, where N_t is the design sample size, and this dependence propagates into the lowest-order terms in the bias. The bias is independent of the test sample size, N_{test}. All measures of classifier performance that count the fraction of times the decision value for an abnormal case exceeds that for a normal case (independent of underlying distribution), and various measures of error for normally distributed decision functions, are nonlinear functions of the parameters of the underlying distributions. They are thus subject to this effect. Fukunaga and Hayes[3] analyzed the finite sample effects on the probability of misclassification (PMC) of a classifier and suggested a technique that makes use of the linear dependence of PMC on $1/N_t$ to estimate the performance at $N_t \rightarrow \infty$ with a finite sample set.

For the evaluation of medical diagnostic systems, the most commonly used performance index is the area under the receiver operating characteristic (ROC) curve, A_z. We have derived analytically that, for linear discriminant classifiers, the classifier performance in terms of A_z can be approximated by a linear function in $1/N_t$, under conditions when higher order terms in N_t can be neglected. We have been investigating the dependence of A_z on sample size by simulation studies.[7–9] Wagner *et al.*[10,11] have also analyzed the effects of design and test sample sizes on the variance components of the classifier performance. Although these behaviors depend strongly on the class distributions and the properties of the classifier, the studies will provide some insight into the sample size requirements for the design of different classifiers. This work may eventually lead to the selection of an efficient resampling scheme for classifier design, as well as the development of a statistical test of the sample size requirements and the generalizability of the trained classifier.

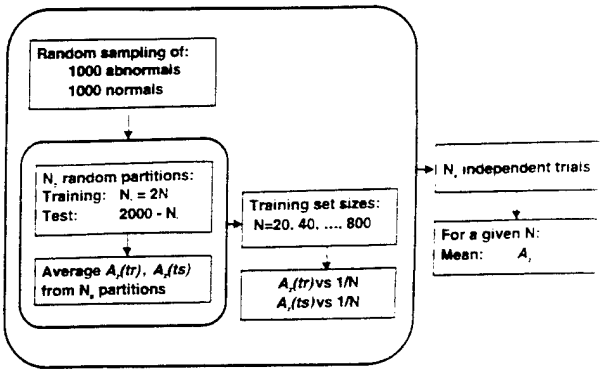


FIG. 1. The sampling and evaluation scheme of the simulation study.

In this paper, we will describe the simulation studies and analyze the effects of sample size on classifier performance. Several commonly used classifiers, including the linear discriminant, the quadratic discriminant, and the back-propagation neural network will be studied and compared under different input conditions. Feature distributions with markedly different characteristics will be used to represent a variety of situations that may be encountered in classification problems for many detection or diagnostic tasks.

II. MATERIALS AND METHODS

We performed simulation studies to evaluate the effects of sample size on classifier design. Normal and abnormal case samples were randomly drawn from known probability distributions of the two classes. These samples were then used to design classifiers for differentiation of normal and abnormal cases. The simulation approach assures that any number of case samples can be obtained from populations with known statistical properties. It thus allows evaluation of the dependence of classifier performance on design sample size and comparison of the performance with theoretically predicted optimal classification based on the chosen probability distributions.

A. Simulation study

The sampling and evaluation scheme of the simulation study is shown in Fig. 1. In this study, we considered only the situation in which equal numbers ($=N_{\text{total}}/2$) of normal and abnormal cases randomly drawn from the class distributions were available in our data set. A resampling strategy similar to the technique suggested by Fukunaga and Hayes was devised to generate the A_z-vs-$1/N_t$ curve. Subsets of $N_{t_1}, N_{t_2}, \ldots, N_{t_j}$ design samples were randomly drawn from the available sample set, again under the constraint that the numbers of normal and abnormal samples were equal in each subset, i.e., $N_{t_i,\text{normal}} = N_{t_i,\text{abnormal}} = N_{t_i}/2$ $(i = 1,\ldots,j)$. A classifier was designed by using each subset of samples. The random sampling of a given subset from the available set of N_{total} samples was performed without replacement, whereas the random sampling of different subsets always started from

the same set of N_{total} samples. Therefore, after drawing a given design subset N_{t_i}, the remaining samples, $N_{\text{total}}-N_{t_i}$ were independent of the design samples and used as the test samples. For simplicity, the number of design samples per class is denoted as N in the following discussion.

In general, there are two methods, resubstitution and hold-out, for testing classifier performance. In the resubstitution method, the design sample set is resubstituted into the trained classifier to test its performance, whereas in the hold-out method, an independent test set is used. It has been shown[1] that, for a Bayes classifier, if the classifier is trained with a finite number of design samples, the resubstitution estimate of the classifier performance is optimistically biased whereas the hold-out estimate is pessimisticaly biased in comparison to that achievable with an infinite design sample set. The mean performance obtained from the former estimation provides an upper bound and that from the latter provides a lower bound on the true classifier performance. When the design sample size is limited, it is important to evaluate the hold-out performance to avoid an overly optimistic prediction of the classifier performance. In the limit of very large sample size, the upper and lower bounds converge towards the unbiased estimate.

In this study, we evaluated the performance of the classifier using both the resubstitution and the hold-out methods as a function of finite design sample size N_t. In order to reduce the variances in the estimates of A_z, we randomly resampled without replacement each N_{t_i} from the same N_{total} samples N_p times, trained and tested the classifier, and estimated the average A_z from the N_p individual A_z's as shown in Fig. 1. The resubstitution or hold-out A_z-vs-$1/N_t$ curve was plotted from the j points and the unbiased estimate of A_z in the limit of $N_t \to \infty$ could be extrapolated from either curve.

This method of estimating classifier performance at large N_t by generating a few data points at finite sample sizes is similar to the Fukunaga and Hayes technique. However, we did not assume that the j points were in the linear region of the A_z-vs-$1/N_t$ curve and we used resampling to reduce the variances. In fact, one of the goals of this study was to investigate the range of design sample size in which the performance curve was approximately linear for various classifiers and probability distributions of the class populations. Therefore, we used a much larger total number of samples ($N_{\text{total}}=2000$) in our simulation study than was generally available for classifier design. We could then choose N_{t_i} over a wide range and study the behavior of the entire A_z-vs-$1/N_t$ curve.

To estimate the population mean of A_z at each N_{t_i}, we repeated the above experiment N_e times, each with 2000 independently drawn samples from the population. The population mean of A_z was estimated by averaging the A_z values obtained from the N_e experiments. We did not analyze the variances in this study because of the complication in the correlation among the N_p values of A_z introduced by resampling. A detailed analysis of the variances and its modeling was performed in a separate study by Wagner *et al.*[10,11] in which a different study design was used.

By varying the number of design samples per class, N, over a large range from 20 to 800, the regime where the $1/N$ dependence dominated could be observed from the A_z (population mean)-vs-$1/N_t$ (or $1/N$) curves. It is important to note that, although the number of test samples, $N_{\text{test}_i}=2000-N_{t_i}$, varied from point to point on both the resubstitution and the hold-out curves, the bias in A_z is independent of N_{test_i}.[1] The shape of the A_z-vs-$1/N$ curve is independent of N_{test_i} after N_{t_i} is fixed. However, the variance of a given A_z does depend on the test sample size.

For simplicity, we will refer to these estimates of A_z (population mean) as $A_z(\text{tr})$ for the resubstitution and as $A_z(\text{ts})$ for the hold-out performance in the following discussions.

B. Class distributions

1. Multivariate normal distributions

For three of the four types of class distributions, we assumed that the normal and abnormal classes followed multivariate normal distributions in the feature space. The dimensionality of the feature space, k, was varied from 3 to 15. The characteristics of the multivariate normal distributions can be completely specified by the multivariate mean vector of the rth class, denoted as $\mu_r\,(r=1,2)$ and its covariance matrix, denoted as Σ_r. The separation of the normal and abnormal classes is measured by the Bhattacharyya distance, B, defined as[1,12]

$$B=\frac{1}{8}\Delta+\frac{1}{2}\ln\frac{\det[(\Sigma_1+\Sigma_2)/2]}{\sqrt{\det\Sigma_1}\sqrt{\det\Sigma_2}}, \tag{1}$$

where $\det\Sigma_r$ denotes the determinant of Σ_r, and Δ is the squared Mahalanobis distance,[12] defined as

$$\Delta=(\mu_2-\mu_1)^T\left(\frac{\Sigma_1+\Sigma_2}{2}\right)^{-1}(\mu_2-\mu_1). \tag{2}$$

The Mahalanobis distance is the Euclidean distance between the means of the two distributions, normalized by the square root of the average of their covariance matrices. It can therefore be considered to be a measure of the signal-to-noise ratio (SNR) between the abnormal and the normal distributions. The second term of B is the contribution from the difference in the covariance matrices of the two class distributions. If the covariance matrices are equal, the second term will be zero and the Bhattacharyya distance will be equal to 1/8 of the squared Mahalanobis distance.

In the current study, three types of multivariate normal class distributions were considered. In the following discussion, we shall refer to the use of simultaneous diagonalization for the two covariance matrices of the class distributions. This operation leaves the normal-based decision functions unchanged because the distance measures that arise in these decision functions are invariant to any non-singular linear transformation.[1]

(1) Equal covariance matrices and unequal means: In this case, the covariance matrices of the normal and abnormal class distributions can be simultaneously diagonalized

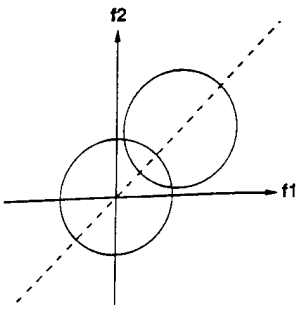


FIG. 2. A schematic illustration of the two class distributions with equal covariance matrices and unequal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

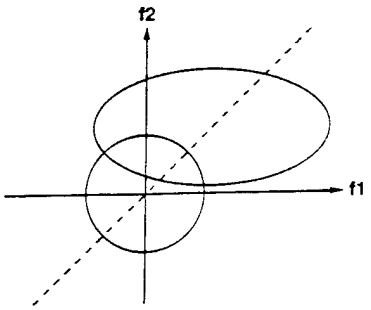


FIG. 3. A schematic illustration of the two class distributions with unequal covariance matrices and unequal means in a 2D feature space. The closed curves represent contours of equal probability in each distribution.

and the variances of the individual feature components can be scaled to unity. Therefore, without loss of generality, the covariance matrices of the two classes could be assumed to be equal to identity matrices, $\Sigma_1=\Sigma_2=I$. The mean feature vector for the first class was assumed to be zero, $\mu_1=\mathbf{0}$, and the mean feature vector for the second class, $\mu_2=\mathbf{M}$ with all components of $\mathbf{M}$ equal to a constant m. The magnitude of m could be adjusted to obtain a desired separation of the two classes. For the purpose of this simulation study, we chose m such that the squared Mahalanobis distance was 3, i.e., the Bhattacharyya distance was 3/8, for feature spaces of any dimensionality. As discussed below, this separation corresponds to a theoretical A_z of 0.89, which is in the performance range of many classification problems in CAD applications. An example of the two class distributions in a 2D feature space is shown schematically in Fig. 2.

(2) Unequal covariance matrices and unequal means: The covariance matrix of the first class was again diagonalized and scaled to be an identity matrix, $\Sigma_1=I$, and the mean feature vector for the first class was assumed to be zero, $\mu_1=\mathbf{0}$. The covariance matrix of the second class, Σ_2, was simultaneously diagonalized to have eigenvalues λ_i, $i=1,...,k$. For this study, we generated the values of λ_i with the simple relationship:

$$\lambda_i=\lambda_{\min}+\frac{(i-1)(\lambda_{\max}-\lambda_{\min})}{(k-1)}, \quad i=1,...,k \tag{3}$$

and evaluated one condition where $\lambda_{\min}=1$, and $\lambda_{\max}=2$ for all dimensionalities of the feature spaces. We also assumed that the components of the mean feature vector μ_2 were equal, the values of which were adjusted to achieve a Bhattacharyya distance of 3/8. For the purpose of demonstrating the general trends of the A_z-vs-$1/N$ curves and comparing the relative performance of the different classifiers under the various conditions, the specific choices of these values are not critical. Figure 3 illustrates an example of the two class distributions in a 2D feature space.

(3) Unequal covariance matrices and equal means: The covariance matrix of the first class was the same as that in the first two cases described above. The covariance matrix of the second class was proportional to the identity matrix, $\Sigma_2=\alpha I$, where the proportionality constant α was adjusted to provide a Bhattacharyya distance of 3/8. The mean feature vectors of the two classes were equal, $\mu_1=\mu_2=\mathbf{0}$. In this case, the discriminatory power of the two classes comes entirely from the difference in the covariance matrices. A schematic of the two class distributions in a 2D feature space is shown in Fig. 4.

2. Checkerboard distributions

The fourth type of class distributions was a checkerboard where the normal and abnormal classes were located in alternate square box regions of the feature space. Within each box of the checkerboard, the feature vectors were uniformly distributed. The two classes did not overlap with each other so that they could be perfectly separated by an "ideal" classifier with $A_z=1$. We considered a 2×3 checkerboard in a 2D feature space and a $2\times2\times2$ checkerboard in a 3D feature space. The example of a 2×3 checkerboard in a 2D feature space is shown in Fig. 5. Such class distributions may not be common in actual classification problems encountered in CAD. However, it was included in this study to demonstrate the capability and limitations of the different classifiers when the class distributions were not multivariate normal.

C. Classifiers

We studied three types of classifiers: the linear discriminants, the quadratic discriminants, and the back-propagation neural networks. They represent a range of classifiers commonly used in the field of pattern recognition at present.

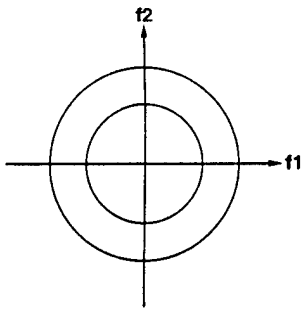


FIG. 4. A schematic illustration of the two class distributions with unequal covariance matrices and equal means in a 2D feature space. The circles represent contours of equal probability in each distribution.

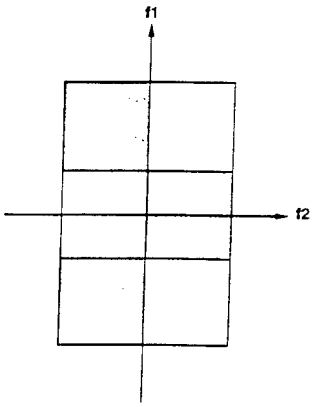


FIG. 5. An example of a 2×3 checkerboard in a 2D feature space.

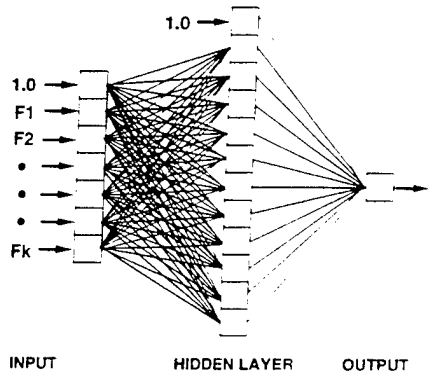


FIG. 6. A schematic diagram of a backpropagation neural network with one hidden layer.

(1) Linear discriminant classifier: The linear discriminant classifier can be derived from the means and the covariance matrices of the class distributions as follows:[1,13]

$$h_l(X)=(\mu_2-\mu_1)^T\bar{\Sigma}^{-1}X+\tfrac{1}{2}(\mu_1^T\bar{\Sigma}^{-1}\mu_1-\mu_2^T\bar{\Sigma}^{-1}\mu_2), \quad (4)$$

where $\bar{\Sigma}=(\Sigma_1+\Sigma_2)/2$, and X is the feature vector to be classified. The means and covariance matrices have to be estimated as the sample means and sample covariance matrices from the available design samples. The sample means and covariance matrices undergo a nonlinear transformation to become the discriminant scores, which in turn are transformed nonlinearly into a measure of the performance. The variances in the estimated parameters propagate into the mean classifier performance and result in a bias through the second derivative of the transformation function.

It is known that, for multivariate normal distributions with equal covariance matrices, the linear discriminant classifier is optimal and the classifier performance in the limit of large design samples is determined by the Mahalanobis distance, given by

$$A_Z=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\sqrt{\Delta/2}}e^{-u^2/2}du. \quad (5)$$

For the class distributions with $\Delta=3$ to be used in this study, it can be derived from Eq. (5) that the maximum A_z that the optimal linear discriminant can achieve in the limit of large design samples is 0.89.

(2) Quadratic discriminant classifier: The quadratic discriminant classifier can be expressed as[1]

$$h_q(X)=\frac{1}{2}(X-\mu_1)^T\Sigma_1^{-1}(X-\mu_1)$$

$$-\frac{1}{2}(X-\mu_2)^T\Sigma_2^{-1}(X-\mu_2)+\frac{1}{2}\ln\frac{\det\Sigma_1}{\det\Sigma_2}. \quad (6)$$

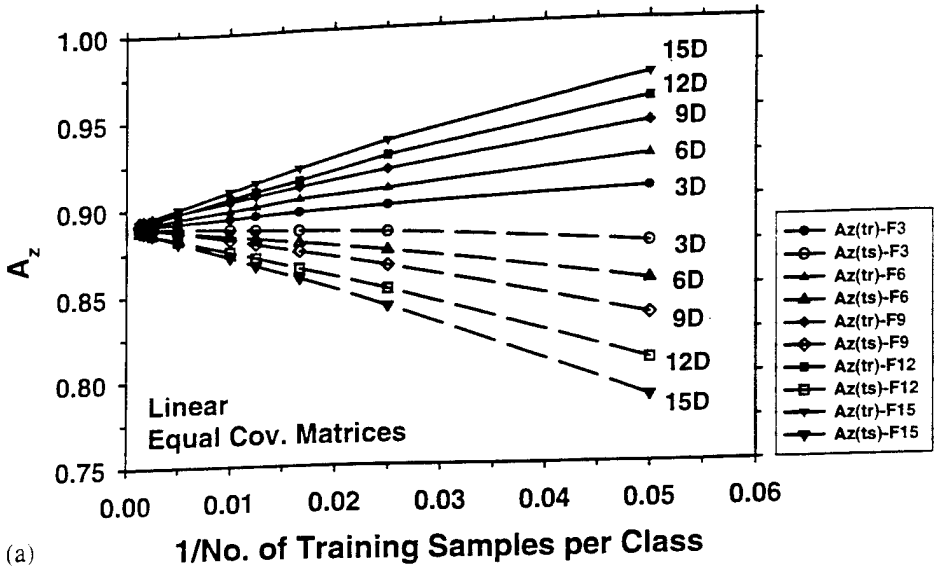
When the class distributions are multivariate normal with unequal covariance matrices, the quadratic discriminant classifier is optimal in the limit of large training samples. The Bhattacharyya distance gives an upper bound on the Bayes error.[1] The general properties of the linear and quadratic classifiers have been described in the literature (for example, Fukunaga[1]).

(3) Back-propagation neural network: Many different architectures and training methods have been developed for artificial neural networks (ANN)[14] in various applications. In this study, we considered only a three-layered neural network trained with a feed-forward back-propagation method. The neural network has k input nodes, n hidden nodes, one output node, and a bias node in both the input and the hidden layers. The ANN architecture is denoted as $k-n-1$. The nodes in the ANN are fully connected and are trained with a minimum sum-of-squares-error criterion. The number of weights to be estimated is equal to $n(k+1)+(n+1)$. A schematic diagram of an ANN is shown in Fig. 6.

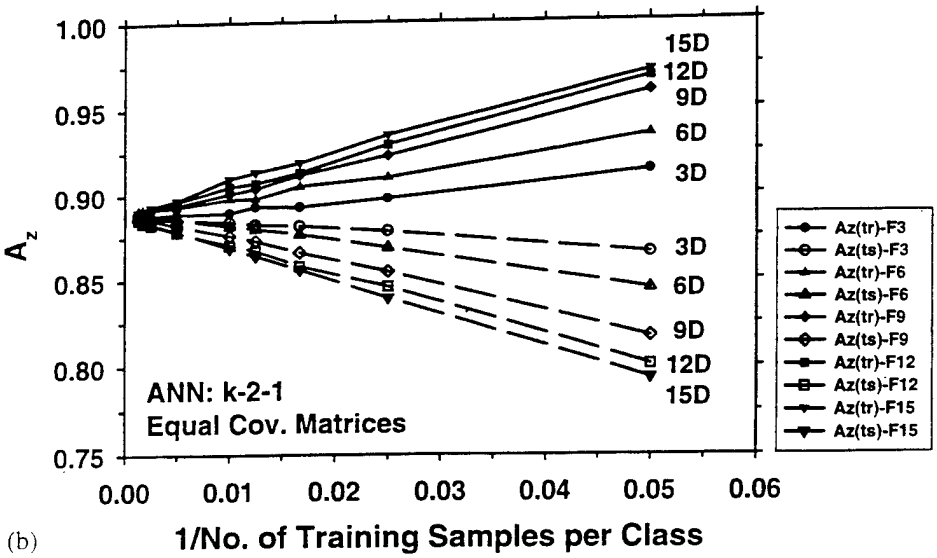
III. RESULTS

In our simulation study, we compared the performance of the linear, quadratic, and backpropagation neural network classifiers for the different class distributions in the feature spaces of dimensionality ranging from 3 to 15. The number of repeated experiments N_e was chosen to be 20 for all cases in the multivariate normal feature spaces and 100 in the checkerboard feature space. The number of data set partitionings N_p in each experiment ranged from 1 to 20. These choices are a compromise between computation time and estimation accuracy, especially for ANN classifiers with a large number of hidden nodes in high dimensional feature spaces. As shown in the graphs discussed below, some of the performance curves may exhibit fluctuations that could be reduced by a larger number of experiments. However, the general trend of the performance curves should not be changed by the statistical uncertainties.

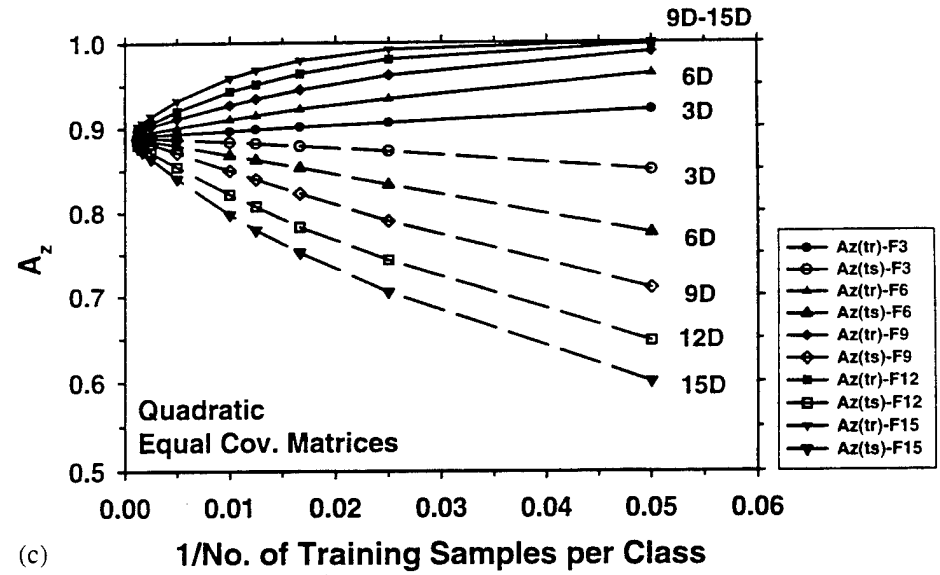
(1) Multivariate normal distributions—Equal covariance matrices and unequal means: For class distributions with equal covariance matrices, the linear discriminant is theoretically the optimal classifier when the design sample size is large. However, when the design sample size is small, the performances of all classifiers are biased. Figures 7(a)–7(c) show the dependence of the A_z obtained from resubstitution (training), A_z(tr), and the A_z obtained from the hold-out method (testing), A_z(ts), on $1/N$ for the linear, ANN, and



(a)



(b)



(c)

FIG. 7. The dependence of the A_z obtained from resubstitution (training-solid lines), A_z(tr), and the A_z obtained from the hold-out method (testing—dashed lines), A_z(ts), on $1/N$ for the class distributions with equal covariance matrices and unequal means. (a) Linear, (b) ANN, and (c) quadratic classifier. Legend: F3=3D feature space, etc.

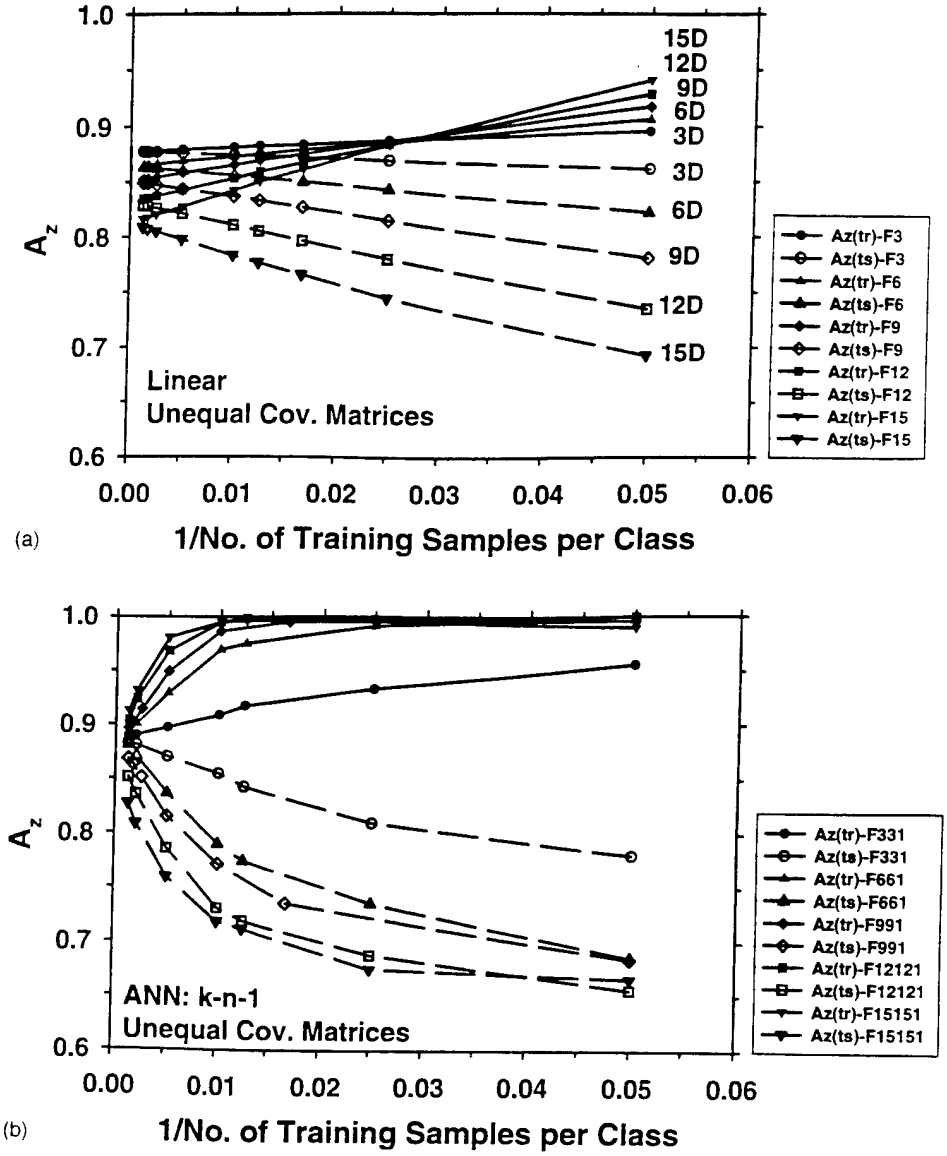


FIG. 8. The performances of the classifiers for class distributions with unequal covariance matrices and unequal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc., solid lines $=A_z$(tr), dashed lines $=A_z$(ts).

quadratic classifier, respectively. Two hidden nodes were used for the ANN $(k-2-1)$ because it is the smallest number of hidden nodes in a nonlinear ANN. An ANN with only one hidden node will be a linear classifier and behave in a similar manner as the linear discriminant. On the other hand, ANNs with a large number of hidden nodes (not shown) will overfit the design samples and have poor generalizability to the unknown cases, similar to the ANN curves to be discussed below. All three classifiers can reach the optimal classification accuracy of $A_z=0.89$ in the limit of large N. The curves for the linear classifier and the ANN $(k-2-1)$ at 400 training epochs (iterations) are approximately linear over the entire range. The quadratic classifier does not reach the approximately linear region until N is greater than about 100 $(1/N<0.01)$ in the higher-dimensional feature space. The biases on both the resubstitution and hold-out curves for the quadratic classifier are greater than those for the linear classifier and the ANN $(k-2-1)$. The large biases again indicate overfitting and poor generalization by the quadratic classifier in the equal-covariance-matrices situation.

(2) Multivariate normal distributions—Unequal covariance matrices and unequal means: The performances of the classifiers for class distributions with unequal covariance matrices are shown in Figs. 8(a)–8(b). The linear discriminant and the ANN $(k-2-1)$ classifier (not shown) are again approximately linear over the entire range of N studied. However, the A_z at $1/N=0$ decreases as the dimensionality of the feature space increases. This is because both the linear discriminant and the near-linear ANN $(k-2-1)$ cannot make use of the class separability due to the differences in the covariance matrices which is the second term in the Bhattacharyya distance. The second term increases relative to the first term, the squared Mahalanobis distance, when the Bhattacharyya distance is fixed and the dimensionality of the feature space increases.

The performance curves of the ANN at large N improve when a greater number of hidden nodes and a sufficient number of training epochs are used. The number of hidden nodes required to reach the optimal classification of $A_z=0.89$ at $1/N=0$ increases with the dimensionality of the feature

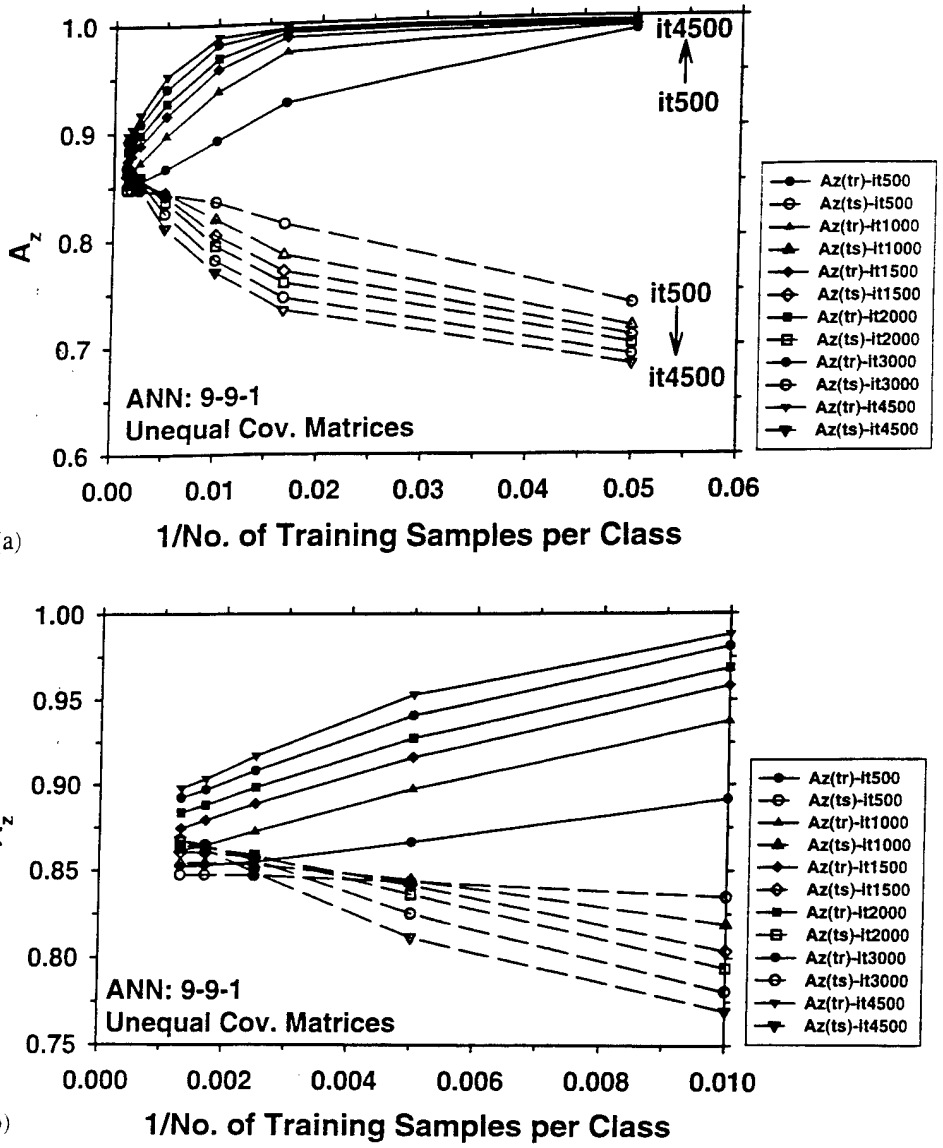


FIG. 9. The dependence of the performance curves on the number of training epochs for an ANN with nine hidden nodes in a 9D feature space: ANN(9−9−1). Legend: it500=500 training epochs, etc., solid lines=A_z(tr), dashed lines=A_z(ts). The expanded view in (b) shows the trend of the curves at large sample sizes.

space. Figure 8(b) shows the performance of the ANNs when the number of hidden nodes is equal to the dimensionality in each feature space. Since the number of weights to be trained increases rapidly with increasing number of nodes in an ANN, the number of epochs required for training the ANN to achieve a reasonable classification accuracy increases accordingly. The resubstitution and hold-out performance curves of each ANN shown in Fig. 8(b) were chosen at the smallest number of training epoch that resulted in approximately the highest A_z value when the hold-out curve was extrapolated to $1/N=0$. The number of training epochs required to reach the highest A_z increased as the dimensionality and the number of hidden nodes in the ANN increased. It ranged from about 4000 to 10 000 for the conditions shown in Fig. 8(b). We did not attempt to perform an exhaustive search for the "optimal" number of hidden nodes in each feature space because of the extensive computation time required for the search. Instead, we evaluated ANNs with a few different numbers of hidden nodes in each feature space and chose the "best" ANN within those studied. With this approximation we observed that, in a k-dimensional feature space and with these class distributions, an ANN with approximately k hidden nodes can approach the optimal performance when the design sample size and the number of training epochs are sufficiently large, as shown in Fig. 8(b).

To illustrate the training of an ANN with a large number of hidden nodes, we show the dependence of the resubstitution and the hold-out curves on the number of training epochs for ANN (9−9−1) in Fig. 9. A number of commonly discussed problems of an ANN can be observed. In the small N region below about 60 samples per class, overparametrization and over-training are obvious, i.e., near perfect classification during training [A_z(tr) greater than 0.95] and poor generalization [A_z(ts) below about 0.8]. The problem becomes more pronounced with an increasing number of training epochs. In the middle range of 200 to 400 samples per class where A_z(ts) increases to a maximum then decreases with further training, an "optimal" number of training epoch exists. Only in the region with a sufficiently large N (greater than about 500 per class), A_z(ts) increases with

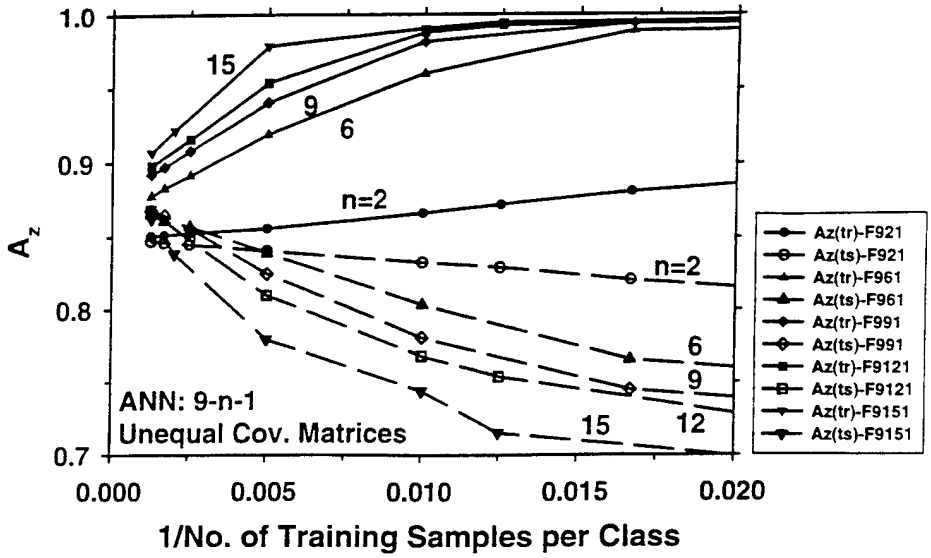


FIG. 10. The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legend: F921 = ANN with two hidden nodes, etc., solid lines = A_z(tr), dashed lines = A_z(ts).

increasing number of training epochs within the range studied. The A_z(ts)-vs-$1/N$ curve becomes linear for N greater than about 200. This dependence of ANN on training epoch is generally observed for ANNs with a large number of hidden nodes and in high-dimensional feature spaces, although the design sample size required in order to avoid overtraining and over-parametrization varies. It reinforces our general experience that the ANNs with a large number of weights can overfit the design samples easily and provide poor generalization when the sample size is small.

The performance curves of ANNs with different numbers of hidden nodes in the 9D feature space are shown in Fig. 10. The curves for a given ANN were again chosen at a training epoch in which the hold-out curve approached approximately the highest performance at $1/N=0$. The chosen training epoch ranged from 600 to 12 000 for the 2- to 15-hidden-node ANNs shown. When the number of hidden nodes is small, the highest A_z obtained by extrapolation to $1/N=0$ appears to be below the theoretical optimum of 0.89. For example, the A_z extrapolated to $1/N=0$ is about 0.85 for ANN (9−2−1), and is about 0.87 for ANN (9−6−1). The ANN with nine hidden nodes appears to approach the optimal A_z of 0.89 in the limit of $1/N=0$. However, the ANN (9−9−1) does not reach the approximately linear region until N is greater than about 200 (easier to see in Fig. 9). As can be seen from the hold-out curves, increasing the number of hidden nodes further will increase overfitting, reduce generalizability, and increase train time without gaining true improvement in performance for classification of unknown case samples.

The quadratic classifier is the theoretically optimal classifier for the class distributions with unequal covariance matrices. It can optimally utilize the class separability contributed by both the differences in the means and the covariance matrices. The performance curves for the quadratic classifier (not shown) in feature spaces of different dimensionalities are very similar to those obtained for the equal covariance matrices situation [Fig. 7(c)]. The A_z of the quadratic classi-

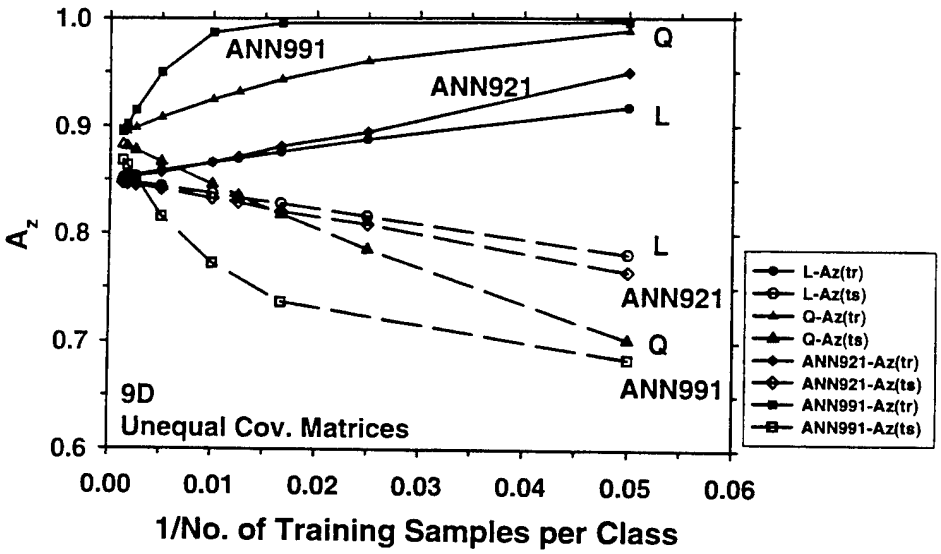


FIG. 11. Comparison of the performance curves of the linear, quadratic, ANN(9−2−1), and ANN(9−9−1) classifiers in the 9D feature space for class distributions with unequal covariance matrices and unequal means. Legends: L=linear; Q=quadratic, ANN=neural network, solid lines = A_z(tr), dashed lines = A_z(ts).

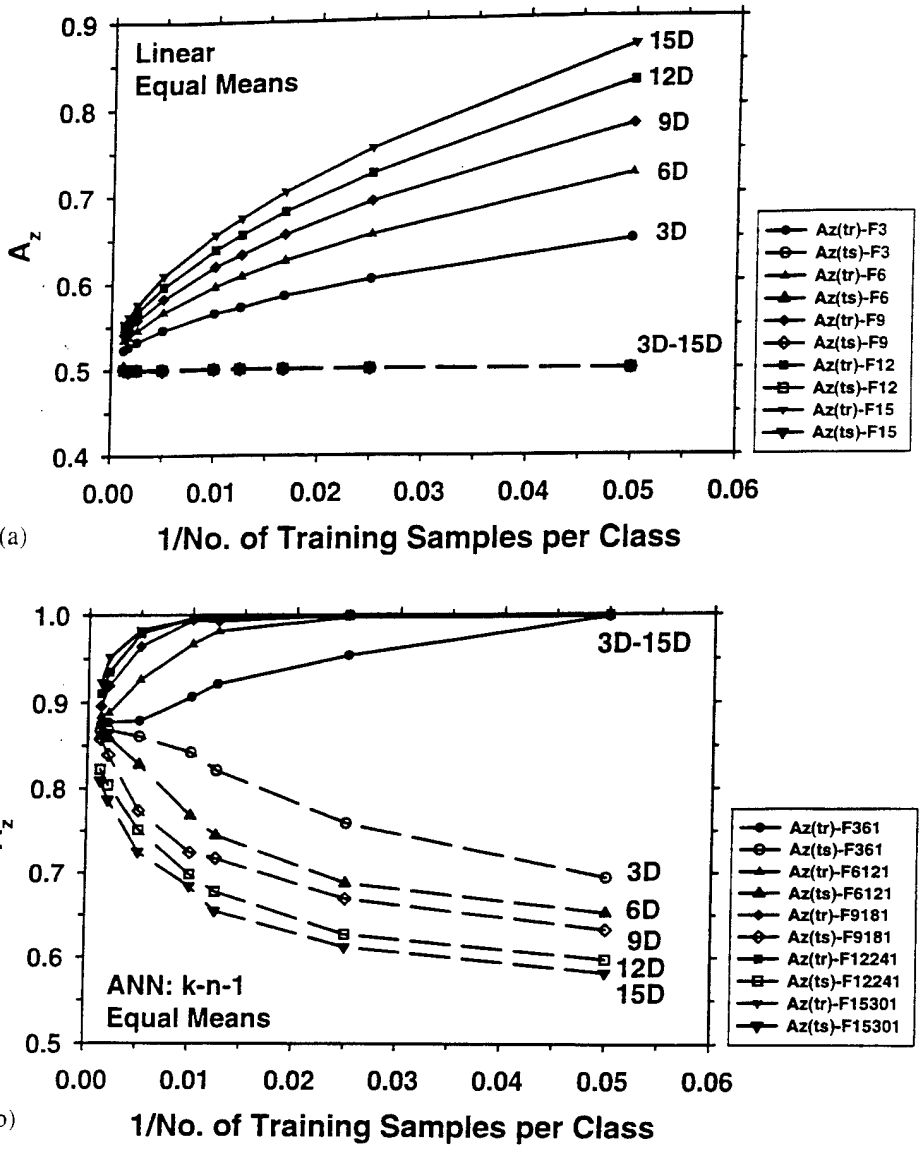


FIG. 12. The dependence of the performance curves on dimensionality of feature space for the class distributions with unequal covariance matrices and equal means. (a) Linear, (b) ANN classifier. Legend: F3=3D feature space, etc. F921 =ANN with two hidden nodes, etc. solid lines=A_z(tr), dashed lines=A_z(ts).

fier reaches the optimal value of 0.89 in the limit of large N for all dimensionalities studied.

Figure 11 shows a comparison of the performance of the linear, quadratic, and the ANN classifiers with two and nine hidden nodes. The biases on the resubstitution and the hold-out curves of the quadratic classifier are not as large as those of the ANN (9−9−1) classifier. However, in the regime of small design sample sizes, the hold-out curve of the optimal quadratic classifier can be much lower than the corresponding curves of the linear classifier or ANN with one or two hidden nodes. This result indicates that the theoretically optimal classifier may not be the optimal choice when the available design sample size is small and overparametrization becomes an important consideration.

(3) Multivariate normal distributions—Unequal covariance matrices and equal means: Figure 12(a) shows the dependence of A_z on $1/N$ for the linear classifiers for the class distributions with equal means. Since the Mahalanobis distance is zero when the means of the two class distributions are equal, the linear classifier performs no better than random guessing in the hold-out situation (A_z(ts)=0.5). However, it is somewhat surprising that the resubstitution curve can be biased to very high A_z values, when the design sample is small. The bias increases with increasing dimensionality of the feature space because the severity of overfitting to the design samples worsens with increased parameterization in the linear discriminant function. This indicates that the predicted performance of a classifier can be unrealistically optimistic if the test samples are not independent of the design samples.

For the class distributions with equal means, it is much more difficult to train the ANN classifier. The number of hidden nodes and the number of training epochs required for the ANN to approximate the decision surfaces, which are spherical hypersurfaces in the k-dimensional feature space, increase as k increases. Figure 12(b) shows the A_z-vs-$1/N$ curves for the ANNs in which the number of hidden nodes is 2 times the dimensionality of the feature space. The number of training epochs required to approach the highest perfor-

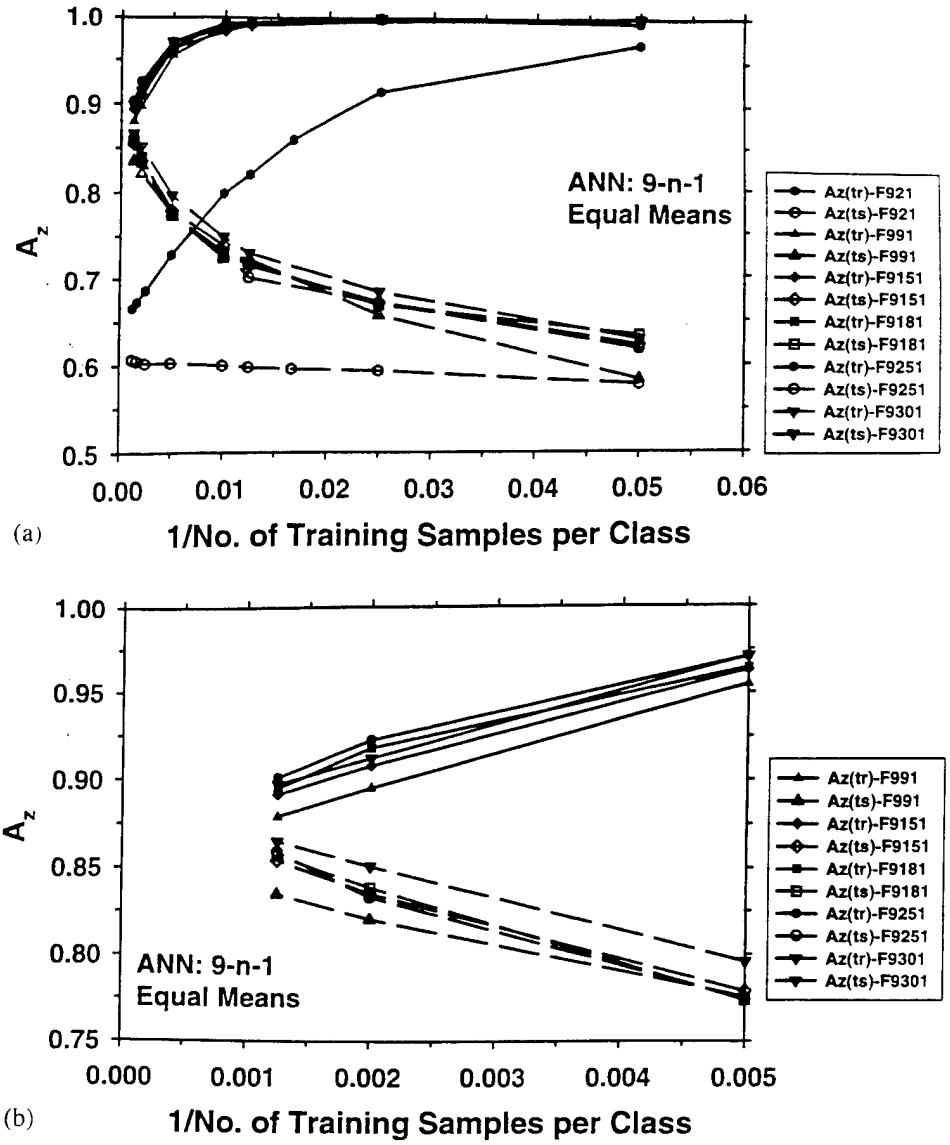


FIG. 13. (a) The dependence of the performance curves of an ANN on the number of hidden nodes in the 9D feature space for class distributions with unequal covariance matrices and equal means. In the expanded scale (b), the approximately linear regions of the curves can be observed. Solid lines $= A_z$(tr), dashed lines $= A_z$(ts).

mance for a given ANN architecture ranges from about 1800 to 20 000 in these cases. Again we did not attempt an exhaustive search for the "optimal" number of hidden nodes in each case. These ANNs were chosen because they appear to approach the maximum performance of $A_z = 0.89$ in the limit of large N and their number of hidden nodes is a simple multiple of the dimensionality. Compared to the class distributions with unequal means, for a given dimensionality, the number of hidden nodes and the number of training epochs required for achieving the near maximum performance at large N are greater in this equal-mean situation. Figure 13(a) shows an example of the dependence of the performance curves on the number of hidden nodes in the 9D feature space. Figure 13(b) is an enlarged view of the curves in Fig. 13(a) in the range where the sample size is greater than 200 per class. The hold-out performance of ANN(9−9−1) at $1/N = 0$ reaches about 0.85. When the number of hidden nodes is greater than nine, the performances of the ANNs at $1/N = 0$ are similar and approach the optimal A_z.

The quadratic discriminant is again the theoretically optimal classifier for the class distributions with unequal covariance matrices. Its performance curves (not shown) are very similar to those plotted in Fig. 7(c), except that the extrapolated A_z values at $1/N = 0$ do not reach as high as those in the equal covariance matrices situation. By using the approximately linear region of the A_z-vs-$1/N$ curve at N greater than 100, the extrapolated A_z ranges from about 0.873 to 0.885 for the 3D to 15D feature spaces. In this case, it is much more efficient to train a quadratic discriminant than the ANN. Since the linear discriminant and ANNs with few hidden nodes cannot provide effective classification regardless of the design sample size, the quadratic discriminant is obviously the optimal classifier both in terms of performance and training efficiency.

(4) Checkerboard distributions: In a feature space with checkerboard class distributions, classification is difficult for many classifiers because of the disjoint clusters of samples belonging to the same class. We compared the three classifiers in such a situation by two examples. Figure 14 shows the performance curves of the three classifiers in a 2D feature

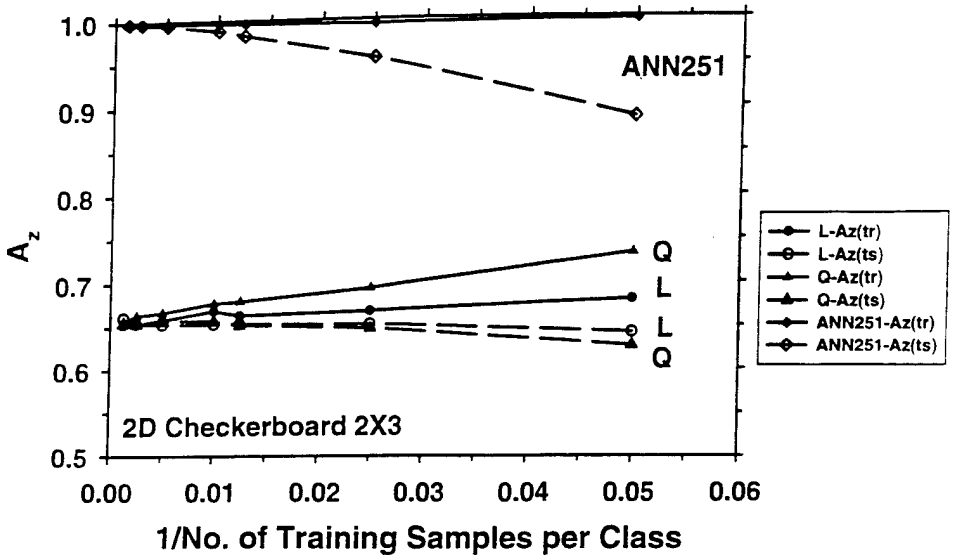


FIG. 14. Performance curves of the three classifiers for a 2×3 unit checkerboard in a 2D feature space. L=linear, Q=quadratic, ANN251=backpropagation neural network with five hidden nodes. Solid lines=A_z(tr), dashed lines=A_z(ts).

space with a 2×3 unit checkerboard distribution. Both the linear and the quadratic discriminants perform poorly even for the resubstitution method where A_z values are in the range of 0.6 to 0.7. However, the ANN(2−3−1) can achieve an A_z of 0.96 (not shown) and the ANN(2−5−1) a near-perfect classification at a training epoch of about 1200.

In a 3D feature space with a 2×2×2 unit checkerboard distribution, the difficulty in classification experienced by the linear and quadratic discriminants is even more apparent. Figure 15 shows that the hold-out curve of the linear classifier is basically the same as random guessing. The hold-out curve of the quadratic classifier is slightly higher than 0.5 at small design sample sizes but approaches 0.5 as the design sample increases. On the other hand, the ANN(3−3−1) can attain a test A_z of 0.9 (not shown) and the ANN(3−5−1) can reach near-perfect classification at large design sample sizes after about 1500 training epochs. These two examples demonstrate that an ANN classifier can be superior to the linear or quadratic classifiers for class distributions that are very different from the idealized multivariate normal distributions.

IV. DISCUSSION

Classifier design is an important field of research in computer-aided diagnosis. Yet many of the issues related to classifier design have not been explored systematically. This simulation study is a part of our on-going investigation of the sample size effects on classifier design.[7–11,15] In this study, we evaluated classifier performance for three multivariate normal class distributions with specific properties: equal covariance matrices, unequal covariance matrices, and equal means. These distributions are idealized but they do approximate a range of situations that may occur in real classification problems. Since the optimal classifier and the upper bound of classification accuracy in the limit of $1/N=0$ are

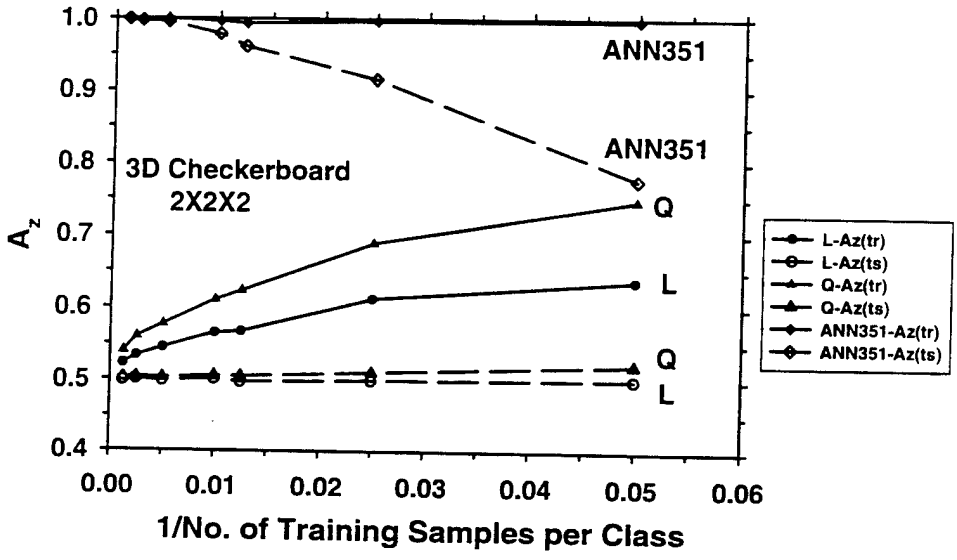


FIG. 15. Performance curves of the three classifiers for a 2×2×2 unit checkerboard distribution in a 3D feature space. Legend: L=linear, Q=quadratic, ANN351=backpropagation neural network with five hidden nodes.

known for each of these cases, we can compare the performances of the classifiers under each condition with the optimum. In addition, a checkerboard class distribution was included in the study. A comparison of the performances of the different classifiers for this class distribution can illustrate their effectiveness when the distributions are very different from multivariate normal.

For all three classifiers, the A_z(tr) obtained by resubstitution is biased optimistically while the A_z(ts) obtained by testing with an independent test set is biased pessimistically, relative to the A_z in the limit of $N \to \infty$, except for the situations when A_z(tr) is bounded from above by perfect classification ($A_z = 1$) or when A_z(ts) is bounded from below by random guessing ($A_z = 0.5$). The magnitude of the biases increases as the design sample size decreases and as the dimensionality of the feature space increases. In the cases where a given classifier has no discriminatory power for a given class distribution, for example, the linear discriminant for the equal-mean or checker-board class distributions, or the quadratic discriminant for the 3D checker-board class distribution, the test A_z(ts) remains almost constant at 0.5, independent of the design sample size. In many cases, the A_z-vs-$1/N$ curve cannot be approximated by a straight line that extrapolates to the A_z at $1/N = 0$ until the design sample sizes are very large, beyond the range of sample sizes that are generally available for CAD classifier design. To estimate the performance of a classifier at large N under the constraint of a small design sample, one may use the Fukunaga and Hayes resampling scheme[3] to derive several points along the A_z-vs-$1/N$ curves in the small sample size region. If the extrapolated resubstitution and hold-out curves do not converge to approximately the same A_z at $1/N = 0$, an average of the points on the two curves which correspond to the same design sample size may be a closer estimate of A_z than either A_z(tr) or A_z(ts). It may be noted that the resubstitution and the hold-out curves are not biased symmetrically from the A_z at infinite N, the average thus obtained will only be a rough estimate. It is also not valid in cases when the classifier has no discriminatory power with A_z(ts) constant at about 0.5 or when the resubstitution curve is overly optimistic with A_z(tr) constant at about 1.

In any case, caution should be taken in estimating classifier performance by extrapolation to $1/N = 0$ or by averaging the resubstitution and hold-out performance as discussed above. The estimated performance contains variances that have to be estimated using further tools. One such attempt in estimating the components of variance by a bootstrapping resampling scheme has been studied recently by Wagner *et al.*[11] These estimates reveal the amount of bias and variance in the classifier performance obtained with the finite design samples, thus allowing estimation of the sample size required to achieve a desired degree of generalizability, rather than replacing the need for a larger sample set and further studies.

With the equal-covariance-matrix class distributions, the linear discriminant is the optimal classifier as expected. The biases are low and the computation is efficient. Moreover, since the A_z-vs-$1/N$ relationship is linear over almost the entire range of design sample sizes, the classifier performance at very large N can be estimated from the small sample size performance by linear interpolation, as suggested by Fukunaga and Hayes[3] and demonstrated previously by Wagner *et al.*[9]

With the unequal-covariance-matrices and equal-mean class distributions, the linear discriminant and the back-propagation neural network with one hidden layer are inferior to the quadratic classifier when the design sample size is large. The linear discriminant cannot utilize the difference in the covariance matrices and underestimates the class separability even when an infinite number of design samples is available. The ANN needs a relatively large number of hidden nodes and a large number of training epochs in order to reach the optimal performance. Its hold-out performance and the computation efficiency are both inferior to those of the quadratic classifier. However, for the unequal-covariance-matrices and unequal-mean case and a small design sample size, the linear classifier or an ANN with very few hidden nodes, e.g., $n = 2$, provides better hold-out performance than the more complex ANNs or the optimal quadratic classifiers. These results indicate that the bias on classifier performance increases with increasing complexity (loosely related to the number of parameters to be estimated) of the classifier. The linear classifier contains $(k+1)$ independent parameters and the quadratic classifier contains $(k+1)(k+2)/2$ independent parameters in their formulations. The number of weights to be estimated for the ANN depends on the number of hidden nodes as $n(k+1)+(n+1)$. The number of weights in an ANN can therefore easily exceed that of a quadratic classifier, although the estimation of the mean and covariance matrices for the linear and quadratic discriminants may contribute additional "complexity" to the classifier design. Two observations can be made. First, when the available sample size is small, a simple classifier will have better generalization than a more complex classifier. Second, a complex ANN or a quadratic classifier trained with an insufficient number of design samples generalizes poorly, even if it is the optimal classifier for the class distributions. It is therefore important to select an appropriate classifier by taking into consideration the design sample size.

A further problem in classifier design is that the true population distributions of the classes in the feature space are generally unknown. It was suggested that the quantile–quantile (Q–Q) plot and the chi-square plot may be used for investigating the normality of univariate and multivariate sample distributions, respectively.[16] However, it is still unknown under what criteria the chi-square plot will indicate that it is optimal to use a classifier designed under the normality assumption. For any measure of goodness-of-fit, when the sample size is small, only the most aberrant deviations from the normal distribution can be identified as a lack of fit from these plots.[16] Therefore, there is often no *a priori* knowledge to select an "optimal" classifier or to predict whether the observed performance is caused by the sample size, the choice of an overly complex classifier, or by an actual poor separation of the classes in the feature space. If one observes poor generalization of a trained classifier in a

truly independent test set, it will be important to take into consideration all these factors and redesign the classifier.

In this study, we assumed that the best features have already been determined for the classification task. In a general classifier design problem, the best set of features usually has to be selected based on the available design samples. The feature selection step will introduce additional biases to the classifier performance. The number of features selected also has a strong influence on the classifier design, as can be seen from the dependence of the bias on the dimensionality of the feature space. The investigation of this more complex situation including both the feature selection and classifier training steps is underway.[17]

The term generalizability is nonspecific and needs to be qualified here. The present paper is concerned with the generalizability of the mean performance of classifiers to unknown test samples drawn from the same population of cases. We have shown in this paper that the mean performance of a classifier depends on the number of samples used to train the classifier, the architecture of the classifier, and—for multivariate-normal data—the means and covariances of the population distributions. Suppose in this context that a classifier is trained on a given finite number of design samples (patients). The mean performance of the classifier over independent replications with the same number of design samples is generalizable to studies characterized by the same number of design samples. In other words, the mean resubstitution or hold-out performance is an unbiased estimate for repeated sampling of independent design and test sample sets, respectively, when the same number of design samples is used. The classifier performance may not, however, be generalizable to studies characterized by a different number of design samples. In particular, when a very large and representative design sample size is used, the mean performance may be very different from the mean performance that characterizes the finite-training-sample condition. When the mean performance under the conditions of a finite design sample size is close to that expected with a very large design sample size, the finite-training sample performance is said to be generalizable to the population performance.

The term generalizability is not only used with respect to mean performance, it is also used with respect to uncertainty in performance, as reflected in estimates of error bars (standard deviations, or the corresponding variances). For example, if we think of repeating a given training and testing experiment on a classifier and if only the test samples are drawn independently on the repeated trials, then the estimated uncertainties are said to be generalizable only to a population of test samples. If, however, we think of repeating the experiment and independently drawing new training samples as well as new test samples, then the estimated uncertainties are said to be generalizable to a population of trainers and a population of testers.[17] Models for the components of variance in both paradigms are the subjects of current work in progress.[10,11] A key point of this latter work is the fact that for computer-aided diagnosis, most available software for ROC analysis only provides estimates of uncertainty that are generalizable to a population of test samples.

In this investigation, we have limited our study to only three types of classifiers: the linear discriminant, the quadratic discriminant, and the backpropagation ANNs with one hidden layer. There are, of course, many other variations of the ANN architecture and other parametric or non-parametric classifiers available for feature classification tasks. The purpose of our work is not to exhaustively evaluate all possible combinations of class distributions and classifiers. Rather, by limiting our investigation to some well-known situations, we can perform systematic analyses and gain some insights into the classifier design problems. Furthermore, we have limited our discussion here to the estimates of the mean classifier performance. Wagner *et al.*[10,11] have investigated the variances of classifier performance estimated from a finite sample set and developed models to study the relative importance of the sizes of the training and test samples. It has been demonstrated that a components-of-variance model can be estimated with a finite sample set by using a bootstrap method. More importantly, the analysis of variances can reveal the generalizability of the performance estimates to other training and test sample sets in the population. Our long term goals are to find some guidelines for designing efficient resampling schemes that can minimize the bias and variance of a trained classifier using the available samples, and to provide a quantitative design tool that can estimate the design sample size requirement for a larger "pivotal" study from the results of a smaller "pilot" study in order to achieve a desired precision in A_z and the desired generalizability.

ACKNOWLEDGMENTS

This work is supported in part by USPHS Grant No. CA 48129 and by a grant from the U.S. Army Medical Research and Materiel Command DAMD 17-96-1-6254, a Career Development Award (B.S.) DAMD 17-96-1-6012 from the U.S. Army Medical Research and Materiel Command and a Whitaker Foundation Grant (N. P.). The content of this paper does not necessarily reflect the position of the government and no official endorsement of any equipment and product of any companies mentioned in this paper should be inferred. The authors are grateful to Charles E. Metz, Ph. D., for providing the LABROC1 programs.

[a]Author to whom correspondence should be addressed. Department of Radiology, University of Michigan, 1500 E. Medical Center Drive, UHB1 F510B, Ann Arbor, MI 48109-0030; Phone: 734-936-4357; Fax: 734-936-7948; Electronic mail: chanhp@umich.edu

[1]K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. (Academic, New York, 1990).

[2]S. Raudys and V. Pikelis, "On dimensionality, sample size, classification error, and complexity of classification algorithm in pattern recognition," IEEE Trans. Pattern. Anal. Mach. Intell. **PAMI-2**, 242–252 (1980).

[3]K. Fukunaga and R. R. Hayes, "Effects of sample size on classifier design," IEEE Trans. Pattern. Anal. Mach. Intell. **11**, 873–885 (1989).

[4]R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, in *Information Processing in Medical Imaging*, edited by H. H. Barrett and A. F. Gmitro (Springer-Verlag, Berlin, 1993).
[5]R. F. Wagner, D. G. Brown, J.-P. Guedon, K. J. Myers, and K. A. Wear, "On combining a few diagnostic tests or features," Proc. SPIE **2167**, 503–512 (1994).
[6]D. G. Brown, A. C. Schneider, M. P. Anderson, and R. F. Wagner, "Effect of finite sample size and correlated/noisy input features on neural network pattern classification," Proc. SPIE **2167**, 180–190 (1994).
[7]H. P. Chan, B. Sahiner, R. F. Wagner, N. Petrick, and J. Mossoba, "Effects of sample size on classifier design: quadratic and neural network classifiers," Proc. SPIE **3034**, 1102–1113 (1997).
[8]H. P. Chan, B. Sahner, R. F. Wagner, and N. Petrick, "Effects of sample size on classifier design for computer-aided diagnosis," Proc. SPIE **3338**, 845–858 (1998).
[9]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Finite-sample effects and resampling plans: applications to linear classifiers in computer-aided diagnosis," Proc. SPIE **3034**, 467–477 (1997).
[10]R. F. Wagner, H. P. Chan, J. T. Mossoba, B. Sahiner, and N. Petrick, "Components of variance in ROC analysis of CADx Classifier performance," Proc. SPIE **3338**, 859–875 (1998).
[11]R. F. Wagner, H. P. Chan, B. Sahiner, N. Petrick, and J. T. Mossoba, "Components of variance in ROC analysis of CADx classifier performance. II: Applications of the bootstrap," Proc. SPIE **3661**, 523–532 (1999).
[12]D. J. Hand, *Discrimination and Classification* (Wiley, New York, 1981).
[13]P. A. Lachenbruch, *Discriminant Analysis* (Hafner, New York, 1975).
[14]J. A. Freeman and D. M. Skapura, *Neural Networks-Algorithms, Applications, and Programming Techniques* (Addison-Wesley, Reading, 1991).
[15]H. P. Chan, B. Sahiner, R. F. Wagner, and N. Petrick, "Classifier design for computer-aided diagnosis in mammography: effects of finite sample size," Med. Phys. **24**, 1034–1035 (1997).
[16]R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis* (Prentice-Hall, Englewood Cliffs, NJ, 1982).
[17]C. A. Roe and C. E. Metz, "Variance-component modeling in the analysis of receiver operating characteristic index estimates," Acad. Radiol. **4**, 587–600 (1997).

Classification of Malignant and Benign Masses Based on Hybrid ART2LDA Approach

Lubomir Hadjiiski,* *Member, IEEE*, Berkman Sahiner, *Member, IEEE*, Heang-Ping Chan, Nicholas Petrick, *Member, IEEE*, and Mark Helvie

Abstract—**A new type of classifier combining an unsupervised and a supervised model was designed and applied to classification of malignant and benign masses on mammograms. The unsupervised model was based on an adaptive resonance theory (ART2) network which clustered the masses into a number of separate classes. The classes were divided into two types: one containing only malignant masses and the other containing a mix of malignant and benign masses. The masses from the malignant classes were classified by ART2. The masses from the mixed classes were input to a supervised linear discriminant classifier (LDA). In this way, some malignant masses were separated and classified by ART2 and the less distinguishable benign and malignant masses were classified by LDA. For the evaluation of classifier performance, 348 regions of interest (ROI's) containing biopsy proven masses (169 benign and 179 malignant) were used. Ten different partitions of training and test groups were randomly generated using an average of 73% of ROI's for training and 27% for testing. Classifier design, including feature selection and weight optimization, was performed with the training group. The test group was kept independent of the training group. The performance of the hybrid classifier was compared to that of an LDA classifier alone and a backpropagation neural network (BPN). Receiver operating characteristics (ROC) analysis was used to evaluate the accuracy of the classifiers. The average area under the ROC curve (A_z) for the hybrid classifier was 0.81 as compared to 0.78 for the LDA and 0.80 for the BPN. The partial areas above a true positive fraction of 0.9 were 0.34, 0.27 and 0.31 for the hybrid, the LDA and the BPN classifier, respectively. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classification in CAD applications.**

Index Terms— **Computer-aided diagnosis, hybrid classifier, mammography, neural networks.**

I. Introduction

MAMMOGRAPHY is the most effective method for detection of early breast cancer [1]. However, the specificity for classification of malignant and benign lesions from mammographic images is relatively low. Clinical studies have shown that the positive predictive value (i.e., ratio of the number of breast cancers found to the total number of biopsies) is only 15% to 30% [2]–[4]. It is important to increase the positive predictive value without reducing the sensitivity of breast cancer detection. Computer-aided diagnosis (CAD) has the potential to increase the diagnostic accuracy by reducing the false-negative rate while increasing the positive predictive values of mammographic abnormalities.

Classifier design is an important step in the development of a CAD system. A classifier has to be able to merge the available input feature information and make a correct evaluation. Commonly used classifiers for CAD include linear discriminants (LDA) [5], [6] and backpropagation neural networks (BPN) [7]–[9] which have been shown to perform well in lesion classification problems [10]–[22]. These classifiers are generally designed by supervised training. However, these types of classifiers have limitations dealing with the nonlinearities in the data (in case of LDA) and in generalizability when a limited number of training samples are available (especially BPN). Another classification approach is based on unsupervised classifiers, which cluster the data into different classes based on the similarities in the properties of the input feature vectors. Therefore, unsupervised classifiers can be used to analyze the similarities within the data. However, it is difficult to use them as a discriminatory classifier [29], [30]. They also have limited generalizability when the training sample set is small.

We propose here a hybrid unsupervised/supervised structure to improve classification performance. The design of this structure was inspired by neural information processing principles such as self organization, decentralization and generalization. It combines the adaptive resonance theory network (ART2) [26], [27] and the LDA classifier as a cascade system (ART2LDA). The self-organizing unsupervised ART2 network automatically decomposes the input samples into classes with different properties. The ART2 network has been found to perform better compared to conventional clustering techniques in terms of learning speed and discriminatory resolution for the detection of rare events in many classification tasks [28]–[30]. The supervised LDA then classifies the samples belonging to a subset of classes that have greater similarities. By improving the homogeneity of the samples, the classifier designed for the subset of classes may be more robust.

The ART2LDA design implements both structural and data decomposition. Decomposition is a powerful approach that can reduce the complexity of a problem. Both structural decom-

Manuscript received January 27, 1999; revised October 26, 1999. This work was supported by in part by the USPHS under Grant No. CA 48129 and in part by the U.S. Army Medical Research and Materiel Command (USAMRMC) under Grant DAMD 17-96-1-6254. The work of L. Hadjiiski was supported in part by the USAMRMC under Career Development Award DAMD 17-98-1-8211. The work of B. Sahiner was supported in part by the USAMRMC under Career Development Award DAMD 17-96-1-6012. The work of Nicholas Petrick was supported in part by a grant from The Whitaker Foundation. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

*L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, and M. Helvie are with the Department of Radiology, The University of Michigan, Ann Arbor, MI 48109-0904 USA.

Publisher Item Identifier S 0278-0062(99)10410-5.

position and data decomposition can improve classification accuracy [23] as well as model accuracy [24]. However, decomposition can also reduce the prediction accuracy due to overfitting the training data. We will demonstrate in this paper that the proposed hybrid structure can reduce the overfitting problem and improve the prediction capabilities of the system. The performance of the hybrid ART2LDA classifier will be compared with those of an LDA alone or a BPN classifier.

The rest of the paper is organized as follows. In Section II the ART2 unsupervised network is described. A hybrid ART2LDA classifier is introduced in Section III. Section IV describes the data set used in this study. The results are presented in Section V. Section VI contains discussion of these results. Finally, Section VII concludes this investigation.

II. ART2 Unsupervised Neural Network

The ART2 is a self-organizing system that can simulate human pattern recognition. ART2 was first described by Grossberg [25] and a series of further improvements were carried out by Carpenter, Grossberg, and coworkers [26]–[28]. The ART2 network clusters the data into different classes based on the properties of the input feature vectors. The members within a class have similar properties. The process of ART2 network learning is a balance between the plasticity and stability dilemma. Plasticity is the ability of the system to discover and remember important new feature patterns. Stability is the ability of the system to remain unchanged when already known feature patterns with noise are input to the system. The balance between plasticity and stability for the ART2 training algorithm allows fast learning [28], i.e., rare events can be memorized with a small number of training iterations without forgetting previous events. The more conventional training algorithms, such as back propagation [7]–[9], perform slow learning, i.e., they tend to average over occurrences of similar events and require many training iterations.

The structure of the ART2 system is shown in Fig. 1. It consists of two parts: the ART2 network and the learning stage. Suppose that there are n input features x_i $(i = 1, \cdots, n)$ and k classes in the ART2 network. When a new vector is presented to the input of the ART2 network, an activation value p_j for class j is calculated as

$$p_j = \sum_{i=1}^{n} x_i w_{ij}, \qquad j = 1, \cdots, k \tag{1}$$

where w_{ij} is the connection weight between input i and class j. The activation value is a measure of the membership of the particular input feature vector to class j. The higher the value p_j is, the better the input vector matches class j. The maximum value p_r is selected from all p_j $(j = 1, \cdots, k)$ to find the best class match. Furthermore, in order to balance the contribution to the activation value from all feature components, the input feature values applied to the ART2 system are scaled between zero and one [30]. This normalization will allow detection of similar feature patterns even when the magnitudes of the input feature components are very different.

The learning stage of the ART2 system can influence the weights of the selected class or the complete ART2 network

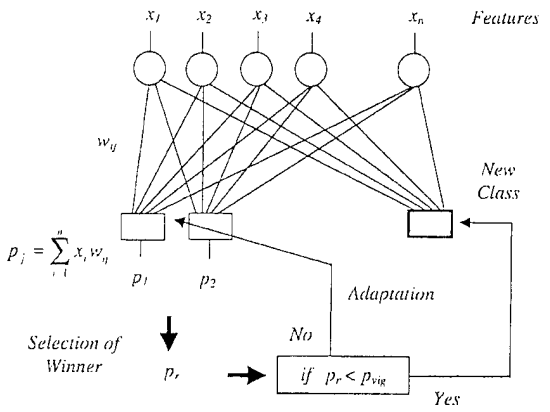


Fig. 1. Structure of the ART2 network.

structure by adding a new class. An additional parameter, the vigilance, is used to determine the type of learning [26]. The vigilance parameter p_{vig} is a threshold value that is compared to the maximum activation value p_r. If p_r is larger than p_{vig} then the input vector is considered to belong to class r. The adaptation of the weights connected with class r is performed as follows:

$$w_{ir}^{new} = w_{ir}^{\text{old}} + \eta(x_i - w_{ir}^{\text{old}}), \qquad \text{for } i = 1, \cdots, n \tag{2}$$

where η is a learning rate. The adaptation of the class r weights (2), aims at maximization of the p_r value for the particular input vector. In an iterative manner the weights are adjusted so that the activation values produced for similar input vectors will be maximum only for the class to which they belong and these maximum activation values will be higher than p_{vig}.

If the maximum activation value p_r is smaller than p_{vig}, it is an indication that a novelty has appeared and a new class will be added to the ART2 structure. The new weights connecting the input with the new class $(k + 1)$ are initialized with the scaled input feature values of this novelty. In such a way, the activation value p_{k+1} will be maximum $(p_r = p_{k+1})$ higher than p_{vig} when computed for this novelty in further training iterations. The value of the vigilance parameter p_{vig} determines the resolution of ART2. It can be chosen in the range between zero and one. In the case that p_{vig} is relatively small, only very different input feature vectors will be distinguished and separated in different classes. If p_{vig} is relatively large, the input feature vectors that are more similar will be separated into different classes. The value of p_{vig} is selected differently depending on the particular application.

III. ART2LDA Classifier

Despite the good performance of ART2 for efficient clustering and detection of novelties, the fast learning approach can cause problems associated with the generalization capability of the system and the correct classification of unknown cases. Supervised classifiers such as linear discriminants or backpropagation neural network classifiers can have better generalization capability than ART2, because they are trained by averaging over similar event occurrences. However, the learning process in these traditional learning algorithms tends

to erase the memory of previous expert knowledge when a new type of expertise is being learned. Therefore, these classifiers do not have as good an ability to correctly classify rare events as ART2 [28], [29].

In order to improve the accuracy and generalization of a classifier, we propose to design a hybrid classifier that combines the unsupervised ART2 network and a supervised LDA classifier. This hybrid classifier (ART2LDA) utilizes the good resolution capability of ART2 and the good generalization capability of LDA. The ART2 first analyzes the similarity of the sample population and identifies a subpopulation that may be separated from the main population. This will improve the performance of the second-stage LDA if the subpopulation causes the sample population to deviate from multivariate normal distributions for which LDA is an optimal classifier. Therefore, the ART2 serves as a screening tool to improve the homogeneity of the sample distributions by classifying outlying samples into separate classes.

The ART2LDA hybrid classifier can be described as

$$y_{AL} = g(f_2(x))f_1(x) + 1 - g(f_2(x)) \tag{3}$$

where x is the input vector, $f_1(\cdot)$ is the LDA classifier, $f_2(\cdot)$ is the ART2 classifier, and $g(\cdot)$ is a binary membership function, which labels the classes identified by ART2 to be one of the two types: malignant class or mixed class. A particular class is defined as malignant if it contains only malignant members. It is defined as mixed if it contains both malignant and benign members. The membership function is defined as follows:

$$g(c) = \begin{cases} 0, & \text{if } c \text{ is a malignant class} \\ 1, & \text{if } c \text{ is a mixed class.} \end{cases} \tag{4}$$

The type of a given class is determined based on ART2 classification of the training data set.

The structure of the ART2LDA classifier is shown in Fig. 2. The ART2 classifies the input sample x into either a malignant or a mixed class. Depending on the class type the function $g(\cdot)$ determines whether the LDA classifier will be used. If x is classified into a mixed class, the final classification will be obtained based on the LDA classifier. However, if x is classified by ART2 into a malignant class, then the mass will be considered malignant, without using the LDA classifier. Therefore, in the ART2LDA structure, the ART2 is used both as a classifier and a supervisor. This can be seen in (3). The first term in (3), $g(f_2(x))f_1(x)$, is the LDA classifier multiplied by the ART2 control part $g(f_2(x))$. The second term in (3), $(1 - g(f_2(x)))$, gives the classification result of the ART2 stage. If $f_2(x)$ is a malignant class, then $g(f_2(x)) = 0$, the LDA stage is eliminated, and the classifier output y_{AL} is equal to 1. On the other hand, if $f_2(x)$ is a mixed class, then $g(f_2(x)) = 1$, the ART2 term is eliminated, and the final classification is determined by the LDA classifier $(y_{AL} = f_1(x))$.

IV. Methods

A. Data Set

The mammograms used in this study were randomly selected from the files of patients who had undergone biopsies at the University of Michigan. The criterion for inclusion of a mammogram in the data set was that the mammogram contained a biopsy-proven mass. The data set contained 348 mammograms with a mixture of benign ($n = 169$) and malignant ($n = 179$) masses. On each mammogram, a region of interest (ROI) containing the mass was identified by a radiologist experienced in breast imaging. The visibility of the masses was rated by the radiologist on a scale of 1 to 10, where the rating of 1 corresponds to the most visible category. The distributions of the visibility rating for both the malignant and benign masses are shown in Fig. 3. The visibility ranged from subtle to obvious for both types of masses. It can be observed that the benign masses tend to be more obvious than the malignant ones. Additionally the likelihood of malignancy for each mass was estimated based on its mammographic appearance. The radiologist rated the likelihood of malignancy on a scale of 1 to 10, where 1 indicated a mass with the most benign appearance. The distribution of the malignancy rating of the masses is shown in Fig. 4.

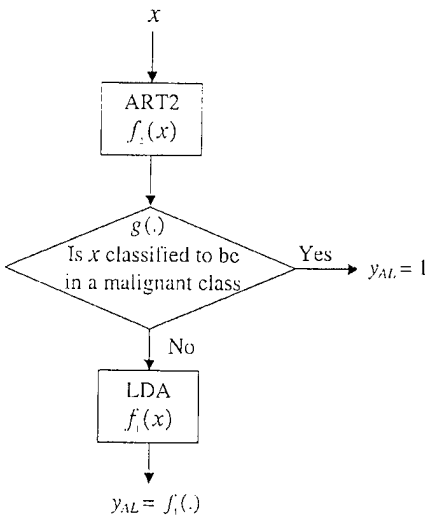


Fig. 2. Structure of the ART2LDA classifier.

The data set can be considered as representative of the patient population that is sent for biopsy under current clinical criteria. Some characteristics of many malignant and benign masses can be visually distinguished by radiologists. However, there is also a nonnegligible fraction of malignant masses that are very similar to benign masses (the low malignancy rating region in Fig. 4). The estimated likelihood of malignancy of malignant and benign masses that are sent for biopsy basically overlaps over the entire range. This is consistent with the fact that in order not to miss malignant masses radiologists must recommend biopsy for even very low suspicion lesions.

Three hundred and five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 μm $\times$ 100 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the optical density (OD) within the range of 0.1 to 2.8 OD units, with a slope of -0.001 OD/pixel value. Outside this range, the slope of the calibration curve decreased gradually. The OD range of the digitizer was 0

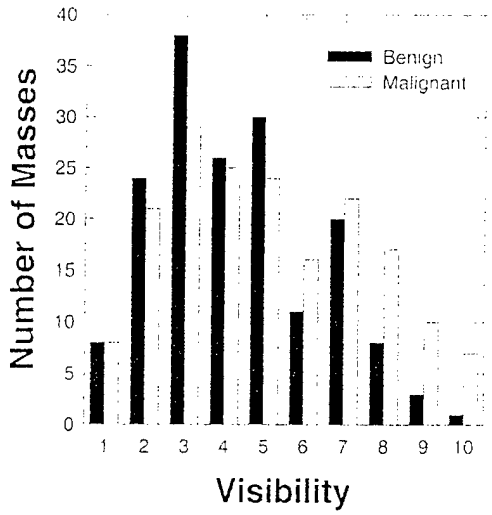


Fig. 3. The distribution of the visibility ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very obvious, 10: very subtle).

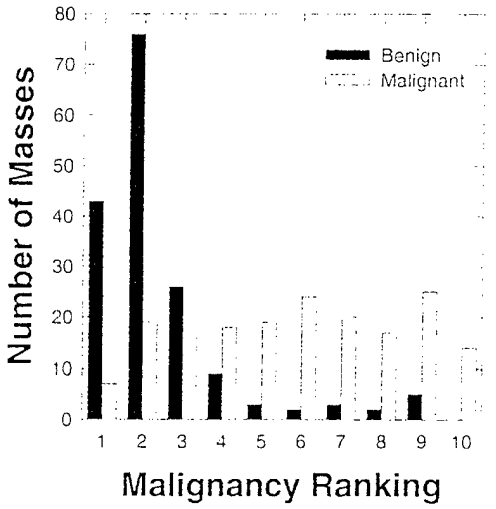


Fig. 4. The distribution of the malignancy ranking of the masses in the dataset. The ranking was performed by an experienced breast radiologist (1: very likely benign, 10: very likely malignant).

to 3.5. The remaining 43 mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel resolution of 50 μm $\times$ 50 μm and 4096 gray levels. The digitizer was calibrated so that gray level values were linearly and inversely proportional to the OD within the range of 0 to 4 OD units, with a slope of -0.001 OD/pixel value. In order to process the mammograms digitized with these two different digitizers, the images digitized with LUMISCAN 85 digitizer were averaged with a 2 $\times$ 2 box filter and subsampled by a factor of two, resulting in 100 μm images.

In order to validate the prediction abilities of the classifier, the data set was partitioned randomly into training and test subsets on a 3:1 ratio, under the constraints that both the malignant and the benign samples were split with the 3:1 ratio and that the images from the same patient were grouped into the same (training or test) subset. These constraints caused the subsets to deviate from an exact 3:1 ratio. The data set was repartitioned randomly ten times. On average, 73% of the samples were grouped into the training set and 27% into the test set. The training and test results from the ten partitions were averaged to reduce their variability.

B. Feature Extraction

A rectangular ROI was defined to include the radiologist-identified mass with an additional surrounding breast tissue region of at least 40 pixels wide from any point of the mass border. A fully automated method was then used for segmentation of the mass from the breast tissue background within the ROI. The rubber band straightening transform (RBST) was previously developed [12] to map a band of pixels surrounding the mass onto the Cartesian plane (a rectangular region). In the transformed image, the border of mass appears approximately as a horizontal edge and spiculations appear approximately as vertical lines. The transformation of the radially oriented textures surrounding the mass margin to a more uniform orientation facilitates the extraction of texture features.

The texture features used in this study were calculated from spatial gray-level dependence (SGLD) matrices [10]–[12], [31], and run-length statistics (RLS) matrices [32] computed from the RBST images. The (i, j)th element of the SGLD matrix is the joint probability that gray levels i and j occur in a direction at a distance of θ pixels apart in an image. Based on our previous studies [10], a bit depth of eight was used in the SGLD matrix construction, i.e., the four least significant bits of the 12-bit pixel values were discarded. Thirteen texture measures, including correlation, energy, difference entropy, inverse difference moment, entropy, sum average, sum entropy, inertia, sum variance, difference average, difference variance, and two types of information measure of correlation were used. These measures were extracted from each SGLD matrix at ten different pixel pair distances ($d = 1, 2, 3, 4, 6, 8, 10, 12, 16$ and 20) and in four directions (0°, 45°, 90°, and 135°). Therefore, a total of 520 SGLD features were calculated for each image. The definitions of the texture measures are given in the literature [10]–[12], [31]. These features contain information about image characteristics such as homogeneity, contrast, and the complexity of the image.

RLS texture features were extracted from the vertical and horizontal gradient magnitude images, which were obtained by filtering the RBST image with horizontally or vertically oriented Sobel filters and computing the absolute gradient value of the filtered image. A gray level run is a set of consecutive, collinear pixels in a given direction which have the same gray level value. The run length is the number of pixels in a run [32]. The RLS matrix describes the run length statistics for each gray level in the image. The (i, j)th element of the RLS matrix is the number of times that the gray level i in the image possesses a run length of j in a given direction. In our previous study, it was found experimentally that a bit depth of five in the RLS matrix computation could provide good texture characteristics [12].

Five texture measures, namely, short run emphasis, long run emphasis, gray level nonuniformity, run length nonuniformity,

and run percentage were extracted from the vertical and horizontal gradient images in two directions, $\theta = 0°$ and $\theta = 90°$. Therefore, a total of 20 RLS features were calculated for each ROI. The formal definition of the RLS feature measures can be found in [32].

A total of 540 features (520 SGLD and 20 RLS) were therefore extracted from each ROI.

C. Feature Selection

In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with stepwise linear discriminant analysis [33] was applied. At each step of the stepwise selection procedure one feature is entered or removed from the feature pool by analyzing its effect on the selection criterion. In this study, the Wilks' lambda (the ratio of within-group sum of squares to the total sum of squares [34]) was used as a selection criterion. The optimization procedure used a threshold F_{in} for feature entry and a threshold F_{out} for feature removal. On a feature entry step, the features not yet selected are entered into the selected feature pool one at a time, the significance of the change in the Wilks' lambda caused by this feature is estimated based on F statistics. The feature with the highest significance is entered into the feature pool if its significance is higher than F_{in}. On a feature removal step, the features which have already been selected are analyzed one at a time from the selected feature pool and the significance of the change in the Wilks' lambda is estimated. The feature with the least significance is removed from the selected feature pool if the significance is less than F_{out}. Since the appropriate values of F_{in} and F_{out} are not known *a priori*, we examined a range of F_{in} and F_{out} values and chose the appropriate thresholds in such a way that a minimum number of features were selected to achieve a high accuracy of classification by LDA for the training sets. More details about the stepwise linear discriminant analysis and its application to CAD can be found in [10]–[12].

D. Performance Analysis

To evaluate the classifier performance, the training and test discriminant scores were analyzed using receiver operating characteristic (ROC) methodology [35]. The discriminant scores of the malignant and benign masses were used as decision variables in the LABROC1 program [36], which fit a binormal ROC curve based on maximum likelihood estimation. The classification accuracy was evaluated as the area under the ROC curve, A_z. For the ART2LDA classifier, the discriminant scores of all case samples classified in the two stages are combined. All masses classified into the malignant group by the ART2 stage were assigned a constant positive discriminant score higher than or equal to the most malignant discriminant score obtained from the LDA stage .

The performance of ART2LDA was also assessed by estimation of the partial area index ($A_z^{(0.9)}$) and compared with the corresponding performance index of the LDA and BPN classifiers. The partial area index ($A_z^{(0.9)}$) is defined as the area that lies under the ROC curve but above a sensitivity threshold of 0.9 ($\text{TPF}_0 = 0.9$) normalized to the total area above TPF_0, (1-TPF_0). The partial $A_z^{(0.9)}$ indicates the performance of the classifier in the high-sensitivity (low false negative) region which is most important for clinical cancer detection task. In addition, the performance of the LDA stage of the ART2LDA classifier was evaluated by the estimation of the area under the ROC curve, denoted as A_z (LDA), for the case samples passed onto the LDA classifier.

TABLE I
NUMBER OF SELECTED FEATURES FOR THE TEN DATA GROUPS WITH THE CORRESPONDING F_{IN} AND F_{OUT} PARAMETERS

Data Group No.	Number of selected features	F_{in}	F_{out}
1	12	1.8	1.6
2	15	2.4	2.2
3	13	2.4	2.2
4	18	2.4	2.2
5	14	2.4	2.2
6	14	2.1	1.8
7	13	2.4	2.2
8	18	1.8	1.6
9	14	2.4	2.2
10	14	2.4	2.2

V. RESULTS

In this section the ART2LDA classification results for malignant and benign masses will be presented and compared with those of the LDA or BPN classifiers. The important point in this study is the fact that the test subset is truly independent of the training subset. Only the training subset is used for feature selection and classifier training, and only the test subset is used for classifier validation. In order to validate the prediction abilities of the classifier, ten different partitions of the training and test sets were used. A different ART2LDA classifier was trained using each training set and the corresponding set of selected features. The classification result was estimated as the average performance for the ten partitions.

For a given partition of training and test sets, feature selection was performed based on the training set alone. The feature selection results for the ten different training groups are shown in Table I. The average number of selected features was 14. An average of two RLS features and twelve SGLD features were selected for each of the training sets which represented 10% of all RLS features and 2.3% of all SGLD features, respectively. Both types of features (RLS and SGLD) are necessary in order to obtain good classification. The most often selected RLS features for the ten training sets were: horizontal short run emphasis (four times), horizontal long run emphasis (six times), vertical run length nonuniformity (three times), horizontal run length nonuniformity (three times). The most often selected SGLD texture measures for the ten training sets were: inverse difference moment (eight times), information measure of correlations one and two (19 times), difference average (nine times), and correlation (ten times). For a given texture measure, features at different angles or distances may be selected, but these features are usually highly correlated so

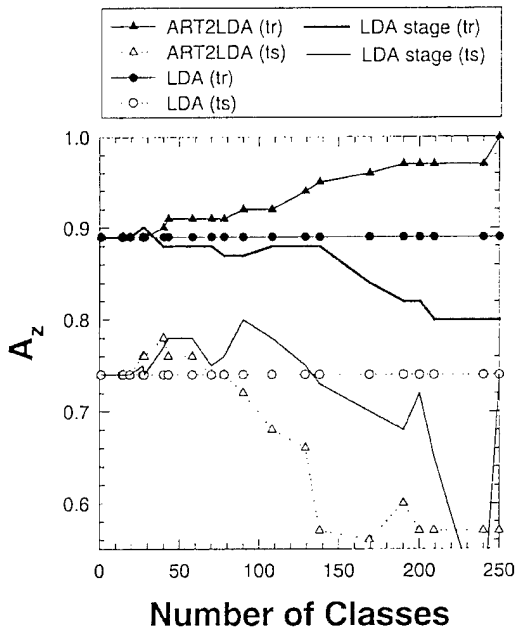


Fig. 5. ART2LDA and LDA classification results for training and test sets from data group three as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

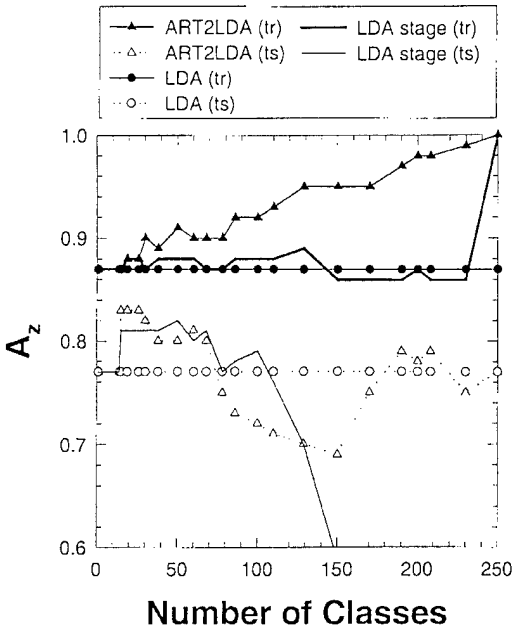


Fig. 6. ART2LDA and LDA classification results for training and test sets from data group one as a function of the generated number of classes. Additionally the results for the LDA stage from the ART2LDA classifier are plotted.

that they can be considered to be similar and counted together as described above.

A. ART2LDA Classification Results

For the ART2LDA classifier, the number of selected features determines the dimensionality of the input vector of the ART2 classifier and the dimensionality of the LDA classifier. By applying different values for the vigilance parameter, ART2 classifiers with different number of classes were obtained. In this study, the vigilance parameter p_{vig} was varied from 0.9 to 0.99, resulting in a range of 10 to 240 classes. The overall performance of the ART2LDA classifier was evaluated for different numbers of ART2 classes because different subset of the samples were separated and classified by ART2 when p_{vig} was varied. In Fig. 5, the classification results for the ART2LDA are compared to the results from LDA alone for the training and test set partition three. The classification accuracy, A_z, was plotted as a function of the number of ART2 classes. For this training and test set partition, when the number of classes was between 20 and 60, the ART2LDA classifier improved the classification accuracy for the test set in comparison to LDA. As the number of classes increased to greater than 60, the A_z value increased for the training data set, but decreased for the test data set and was lower than that of the LDA alone. The two solid lines in Fig. 5 show the A_z values for the LDA stage in the ART2LDA classifier for both the training and test sets. It can be observed that the test A_z for the LDA stage is higher than the A_z for the LDA classifier alone, but not as high as A_z obtained by ART2LDA when the number of classes is small.

In Fig. 6 the classification results of LDA and ART2LDA for the partition one training and test sets are shown. In this case it appeared that in the test set there were two large malignant outliers which degraded the LDA performance. Only 15 classes at the ART2 stage in the ART2LDA was enough to cluster the outliers into a separate malignant class and to improve the performance of the LDA stage and the overall result. The rest of the outliers required more ART2 classes before they were clustered into separate classes and correctly classified as malignant. This is the reason for the similar behavior of the classifiers for partitions three and one in the range of 40 to 70 classes as seen in Figs. 5 and 6. When the number of classes was less than 70, the test A_z for the LDA stage (A_z(LDA)) was higher than the LDA alone, but not as high as the A_z for ART2LDA with less than 30 classes (Fig. 6). The best A_z values for the test data sets of the ten training and test partitions are presented in Table II and Fig. 7. The ART2LDA classifier achieved higher A_z values than the LDA alone in nine of the ten partitions. The average A_z is 0.81 for ART2LDA and 0.78 for LDA alone. The standard deviations of the A_z values for the ten groups range from 0.03 to 0.05 for the ART2LDA classifier and from 0.04 to 0.05 for the LDA classifier.

The performance of ART2LDA was also assessed by estimation of the partial area under the ROC curve $A_z^{(0.9)}$ at a TPF higher than 0.9. The results are presented in Table III and Fig. 7. In the lower part of Fig. 7, the $A_z^{(0.9)}$ values of the test set for the corresponding ten partitions of training and test sets are presented. The average test $A_z^{(0.9)}$ value is 0.34 for the ART2LDA and 0.27 for LDA. For nine of the ten partitions, the $A_z^{(0.9)}$ value was improved at the high-sensitivity operating region (TPF $>$ 0.9) of the ROC curve.

The classifier performance was also evaluated when the ART2LDA classifiers were designed using a fixed number

TABLE II
CLASSIFIERS PERFORMANCE FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE TOTAL AREA UNDER ROC CURVE

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.77	0.83	0.85	0.80
2	0.78	0.80	0.82	0.77
3	0.74	0.78	0.77	0.78
4	0.77	0.77	0.75	0.77
5	0.77	0.78	0.76	0.77
6	0.80	0.83	0.82	0.81
7	0.80	0.81	0.82	0.77
8	0.77	0.80	0.74	0.75
9	0.77	0.80	0.81	0.80
10	0.86	0.89	0.84	0.89
Mean	0.78	0.81	0.80	0.79

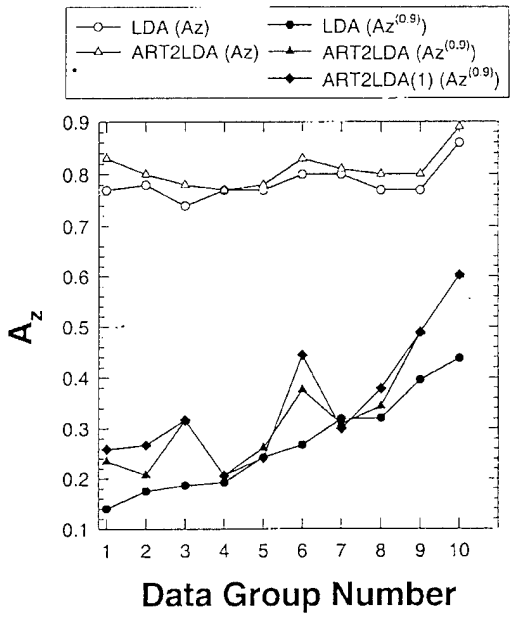


Fig. 7. Average A_z classification results for the 10 test sets. The top graphs represent the ART2LDA and LDA A_z values for the total area under the ROC curve. The bottom graphs represent the ART2LDA, ART2LDA(1) and LDA A_z values for the partial area of the ROC curve above the true positive fraction of 0.9.

TABLE III
CLASSIFIERS RESULTS FOR THE TEN TEST SETS. THE A_z VALUES REPRESENT THE PARTIAL AREA OF THE ROC CURVE ABOVE THE TRUE POSITIVE FRACTION OF 0.9 ($A_z^{(0.9)}$)

Data Group No.	LDA	ART2LDA	BPN	ART2LDA(1)
1	0.14	0.23	0.31	0.26
2	0.17	0.21	0.28	0.27
3	0.19	0.32	0.27	0.32
4	0.19	0.21	0.19	0.21
5	0.24	0.26	0.32	0.24
6	0.27	0.38	0.27	0.44
7	0.32	0.31	0.38	0.30
8	0.32	0.34	0.25	0.38
9	0.40	0.49	0.40	0.49
10	0.44	0.60	0.38	0.60
Mean	0.27	0.34	0.31	0.35

of ART2 classes. The A_z, and $A_z^{(0.9)}$ results, averaged over the ten test partitions, are presented in Table IV. The average A_z with the ART2LDA classifier, compared to that of LDA alone, was again improved between 15 and 40 classes. The maximum average A_z of 0.80 was achieved between 20 and 40 classes. The average $A_z^{(0.9)}$ results are improved for all

TABLE IV
AVERAGE A_z AND AVERAGE $A_z^{(0.9)}$ CLASSIFICATION RESULTS FOR THE TEN TEST SETS. CLASSIFIERS WERE DESIGNED USING A FIXED NUMBER OF ART2 CLASSES

	LDA	ART2LDA					
No. of classes		15	20	30	40	50	60
A_z	0.78	0.80	0.80	0.80	0.80	0.78	0.77
$A_z^{(0.9)}$	0.27	0.30	0.31	0.33	0.33	0.31	0.31

ART2LDA classifiers presented in Table IV. The maximum average $A_z^{(0.9)}$ value is 0.33 and it remains constant between 30 and 40 classes.

An alternative way to evaluate the performance of a classifier is its classification accuracy when a decision threshold for malignancy is selected based on the training set. For instance, a decision threshold may be selected such that all positive samples from the training set are classified correctly i.e., at a sensitivity of 100%. The ART2LDA with this decision threshold is referred to as ART2LDA(1). For a given training and test partitioning, ART2LDA classifiers with different number of classes in the ART2 stage were obtained (Figs. 5 and 6). For each of these models the decision threshold for a sensitivity of 100% was selected from the training set and the corresponding ART2LDA(1) classifier was obtained. Then the ART2LDA(1) classifier (with a specific number of classes in the ART2 stage) that correctly classified the maximum number of malignant masses in the test set is selected. By using all samples of the test set, the A_z value is calculated for the corresponding ART2LDA model. The A_z values for the ART2LDA(1) classifiers for the test sets of the ten data partitionings are shown in Tables II and III. For five of the partitions the overall A_z value for ART2LDA(1) is higher than that of LDA alone (Table II). The average A_z value was 0.79. The partial areas above the TP fraction of 0.9, $A_z^{(0.9)}$, for the ten test data sets obtained by the ART2LDA(1) classifier are also shown in Fig. 7. The ART2LDA(1) achieved the highest average $A_z^{(0.9)}$ value of 0.35 compared to ART2LDA and LDA (Table III).

B. BPN Classification Results

A multilayer perceptron back-propagation neural network with a single hidden layer and a single output node was used for comparison with the ART2LDA classifier. The number of selected features determined the number of input nodes to the BPN. The same ten training/test set partitions (as in the case of ART2LDA) were used for the training and validation of the BPN classifiers. BPN's with their number of hidden nodes ranging from two to ten were evaluated to obtain the best architecture. Back-propagation training was used. Each of the BPN's was trained for up to 18 000 training epochs. At every 1000 epochs the neural network weights were saved and the classification result for the corresponding test set was evaluated. This design procedure was repeated for each of the ten training/test groups. For each group, the best test result among all the BPN architectures (different number of hidden nodes) and all the training epochs examined was selected. The average test A_z over the ten groups for the BPN was 0.80, compared to 0.81 for ART2LDA (Table II). The standard deviations of the A_z values for the ten groups range from 0.04 to 0.05 for the BPN. The average partial $A_z^{(0.9)}$ for the BPN

was 0.31, compared to 0.34 for ART2LDA (Table III). The A_z and $A_z^{(0.9)}$ of the ART2LDA classifier were higher than those of the BPN in six of the ten training/test groups.

VI. Discussion

In the present study, a new classifier (ART2LDA) was designed and applied to the classification of malignant and benign masses. The results indicated that the ART2LDA classifier had better generalizability than an LDA classifier alone. The ART2 classifier grouped the case samples that were different from the main population into separate classes. The minimum number of classes needed to start the clustering of outliers into separate classes depended on how different the outliers were from the rest of the sample population. For the ten different partitions of training and test sets used in this study, the minimum number varied between 13 and 15 classes. When the number of ART2 classes was less than this minimum number of classes, the ART2 classifier generated only mixed malignant-benign classes and all samples were transferred to the LDA stage. In that case, the ART2LDA was equivalent to the LDA classifier alone. When a higher number of classes were generated, an increased number of cases that might be considered outliers of the general data population was removed (clustered in separate classes). For the ten training sets used in this study, the malignant outliers were gradually removed when the number of classes increased. The training accuracy increased when the number of classes increased and A_z could reach the value of 1.0. However, a large number of ART2 classes led to overfitting the training sample set and poor generalization in the test set. The classification accuracy of ART2 for the test set tended to decrease when the number of classes was greater than about 70. The large number of classes also led to a reduction in the generalizability of the second-stage LDA; the training of LDA with a small number of samples would again result in overfitting the training set, and poor generalizability in the test set. This effect was observed when more than 60 or 70 classes were generated by ART2 (see Figs. 5 and 6).

The classification accuracy of ART2LDA increased initially with an increased number of classes and then decreased after reaching a maximum. The correct classification of the outliers by the ART2 in combination with an improvement in the classification by the LDA resulted in the increased accuracy. When the number of ART2 classes was further increased, the effects of overfitting by the ART2 and the LDA became dominant and the prediction ability of the ART2LDA decreased. In some cases the second-stage LDA prediction was much worse than the ART2. In other cases the ART2 could not generalize well. The generation of a high number of classes is therefore impractical and unnecessary both from a computational and a methodological point of view.

For the optimal number of classes (usually less than 50 for the data sets used) the A_z value for the second-stage LDA in the ART2LDA was better than an LDA classifier alone, but it was not as good as the overall A_z from the ART2LDA. It is evident that the ART2 was a useful classifier for improvement of the second-stage classification.

When the partial area of the ROC curve above the true positive fraction (TPF) of 0.9 ($A_z^{(0.9)}$) was considered as a measure of classification accuracy, the advantage of ART2LDA over LDA alone became even more evident. By removing and correctly classifying the outliers, the accuracy of the classification was increased at the high sensitivity end of the curve.

The classifier performance was evaluated when the ART2LDA classifiers were designed using a fixed number of ART2 classes. The results showed improved performance of the ART2LDA in a range between 20 and 40 ART2 classes. Both the average A_z and the average $A_z^{(0.9)}$ reached a maximum within this region, and the maximum average A_z and the average $A_z^{(0.9)}$ values remained unchanged between 30 and 40 classes. These results indicated that the performance of a hybrid ART2LDA classifer was robust and stable and could be potentially useful in real clinical applications.

We have performed statistical tests with the CLABROC program to estimate the significance in the differences between the A_z values from the ART2LDA, the LDA alone, and the BPN, as well as in the differences in the partial $A_z^{(0.9)}$ from the three classifiers. The statistical tests were performed for each individual data set partition because the correlation among the data sets from the different partitions precludes the use of student's paired t test with the ten partitions. We found that the differences in both cases did not reach statistical significance because of the small number of test samples and thus the large standard deviation in the A_z values. However, the consistent improvements in A_z and $A_z^{(0.9)}$ by the ART2LDA (9 out of 10 data set partitions in both cases for LDA and six out of ten data set partitions in both cases for BPN) suggest that the improvement was not by chance alone, and that the accuracy of a classification task could be improved by the use of an ART2 network. In addition, one advantage of the ART2LDA is that the training process is more efficient than that of the BPN, especially when there is a subset of outlying samples. In such a case, the BPN will require a large number of training epochs to minimize the error function.

ART2LDA can be trained to classify the sample cases into more than two classes, such as a class of normal tissue regions in addition to malignant and benign masses. There will be an increase in the complexity of training and a larger training sample size will be desired, but these requirements will be comparable for the different classifiers. In a clinical situation, if the classification task is performed on all computer-detected lesions, the classifier has to distinguish the falsely detected normal tissue from malignant or benign lesions. However, it may be noted that a classifier that can distinguish only malignant and benign masses is applicable to the scenario that the radiologist identifies a suspicious lesion on the mammogram and would like to have a second opinion about its likelihood of malignancy before making a diagnostic decision. Therefore, the development of a classifier that can differentiate malignant and benign masses is the research of interest for many investigators.

Similarly, ART2 can be trained to discover and remove a pure benign mass class. The approach will be similar to the task of classifying and removing the pure malignant classes,

as described in this study. However, our approach of removing the malignant classes will reduce the chance of misclassification of malignant masses. In breast cancer detection, the cost of false-negative (missed cancer) is very high. Therefore, our goal in classifier design is to be conservative. By removing the malignant classes in the first stage, any misclassification to these classes will be regarded as malignant. The remaining classes will be classified again with the second-stage classifier so malignant masses will be less likely to be missed.

The problem of classification of malignant and benign masses has been studied by many investigators. Rangayyan *et al.* [15] used Mahalanobis distance classifer (a modification of an LDA classifier) and the leave-one-out method to evaluate the classification of 54 masses. Fogel *et al.* [16] compared LDA and BPN classifiers using the leave-one-out method and 139 masses (malignant and benign classification). Highnam *et al.* [17] used a morphological feature called a halo to classify 40 masses as malignant and benign. Huo *et al.* [22] employed BPN and a rule-based classifier to classify 95 masses using the leave-one-out evaluation method. Sahiner *et al.* [12] used an LDA classifier and the leave-one-out method to classify 168 masses. An important difference between the classifier designed in this study and the previous studies in the CAD field is the method of feature selection. In the above mentioned studies [12], [15]–[17], [22] and several other published studies [18]–[21] the features were selected from the entire data set first, and then the data set was partitioned into training and test sets. This meant that at the feature selection stage of the classifier design, the entire data set was used as a training set. Depending on the distribution of the features and the total number of samples used, the test results in these studies might be optimistically biased [37]. In our current study, the entire data set was initially partitioned into training and test sets and then feature selection was performed only on the training set. This method will result in a pessimistic estimate of the classifier performance when the training set is small [37]. However, it will provide a more conservative but realistic estimation of the classifier performance in the general patient population. We can expect that the performance would be improved if the classifier in this study were designed using a large data set. Since our main purpose in this study was to compare the ART2LDA classifier with the commonly used LDA and BPN, we did not attempt to quantify how pessimistic our results were in this study.

The most important contribution of this paper is to introduce a new approach that utilizes a two-stage unsupervised–supervised hybrid classifier. We believe that the hybrid approach will improve classification when the sample distribution contains subpopulations that may be difficult for a single classifier to classify. It will be useful for similar classification tasks although different classifiers may be used in each stage of the hybrid structure.

VII. Conclusion

A new classifier combining an unsupervised ART2 and a supervised LDA has been designed and applied to the classification of malignant and benign masses. A data set consisting of 348 films (179 malignant and 169 benign) was randomly partitioned into training and test subsets. Ten different random partitions were generated. For each training set, texture features were extracted and feature selection was performed. An average of features were selected for each group. A hybrid ART2LDA classifier, an LDA, and a BPN were trained by using each of the ten training sets. The A_z value under the ROC curve for the test sets, averaged over the ten partitions, was higher for ART2LDA ($A_z = 0.81$) compared to those of the LDA alone ($A_z = 0.78$) and of the BPN ($A_z = 0.80$). A greater improvement was obtained when the partial ROC area above a true-positive fraction of 0.9 was considered. The average partial A_z for ART2LDA was 0.34, as compared to 0.27 for LDA and 0.31 for BPN. Additionally, for the ART2LDA classifiers that correctly classified the maximum number of malignant masses in the test sets with decision threshold defined with the training set, the average partial A_z was 0.35. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

Acknowledgment

The authors would like to thank Prof. S. Grosberg and Dr. G. Carpenter for providing them with valuable information as well as for the useful discussions. Additionally the authors would like to thank C. E. Metz, Ph.D., for providing the LABROC1 and CLABROC programs.

References

[1] H. C. Zuckerman, "The role of mammography in the diagnosis of breast canser," in *Breast Canser, Diagnosis and Treatment*, I. M. Ariel and J. B. Cleary, Eds. New York: McGraw-Hill, 1987, pp. 152–172.
[2] D. B. Kopans, "The positive predictive value of mammography," *Amer. J. Roentgenol.*, vol. 158, pp. 521–526, 1992.
[3] D. D. Adler and M. A. Helvie, "Mammographic biopsy recommendations," *Curr. Opin. Radiol.*, vol. 4, pp. 123–129, 1992.
[4] M. Moskowitz, "Impact of a priory medical detection on screening for breast cancer," *Radiology*, vol. 184, pp. 619–622, 1989.
[5] P. A. Lachenbruch, *Discriminant Analysis*. New York: Hafner, 1975.
[6] R. O. Duda, and P.E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley, 1973.
[7] P. J. Werbos, "Beyond regression: New tools for prediction and analysis in the behavioral sciences," Ph.D. dissertation, Harvard Univ., Cambridge, MA, 1974.
[8] D. Rumelhart, G. E. Hinton, and R. J. Williams, in D. E. Rumelhart, Ed., *Parallel and Distributed Processing*. Cambridge, MA: MIT Press, 1986, vol. 1, p. 318.
[9] J. Herz, A. Krogh, and R. Palmer, *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley, 1991.
[10] H. P. Chan, D. Wei, M. A. Helvie, B. Sahiner, D. D. Adler, M. M. Goodsitt, and N. Petrick, "Computer-aided classification of mammographic masses and normal tissue: Linear discriminat analysis in texture feature space," *Phys. Med. Biol.*, vol. 40, pp. 857–876, 1995.
[11] D. Wei, H. P. Chan, M. A. Helvie, B. Sahiner, N. Petrick, D. D. Adler, and M. M. Goodsitt, "Classification of mass and normal breast tissue on digital mammograms: Multiresolution texture analysis," *Med. Phys.*, vol. 22, pp. 1501–1513, 1995.
[12] B. Sahiner, H. P. Chan, N. Petick, M. A. Helvie, and M. M. Goodsitt, "Computerized characterization of masses on mamograms: The rubber band sraightening transform and texture analysis," *Med. Phys.*, vol. 25, no. 4, pp. 516–526, Apr. 1998.
[13] B. Sahiner, H. P. Chan, D. Wei, N. Petick, M. A. Helvie, D. D. Adler, and M. M. Goodsitt, "Image feature selection by a genetic algorithm: Application to classification of mass and normal breast tissue," *Med. Phys.*, vol. 23, no. 10, pp. 1671–1683, Oct. 1996.
[14] H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, K. L. Lam, D. D. Adler, and M. M. Goodsitt, "Computerized classification of malignant

and benign microcalsifications on mammograms: Texture analysis using an artificial neural network," *Phys. Med. Biol.*, vol. 42, pp. 549–567, 1997.

[15] R. M. Rangayyan, N. M. El-Farmawy, J. E. Desautels, and O. A. Alim, "Measures of acutance and shape for classification of breast tumors," *IEEE Trans. Med. Imag.*, vol. 16, pp. 799–810, Dec. 1997.

[16] D. B. Fogel, E. C. Wasson, E. M. Boughton, V. W. Porto, and P. J. "Angeline, linear and neural model for classifying breast masses," *IEEE Trans. Med. Imag.*, vol. 17, pp. 485–488, June 1998.

[17] R. P. Highnam, J. M. Brady, and B. J.Shepstone, "A quantitative feature to aid diagnosis in mammography," in *Proc. Digital Mammography'96*, pp. 201–206.

[18] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: Application to decision making in the diagnosis of breast cancer," *Radiology*, vol. 187, pp. 81–87, 1993.

[19] V. Goldberg, A. Manduca, D. L. Evert, J. J. Gisvold, and J. F. Greenleaf, "Improvements in specificity of ultrasonography for diagnosis of breast tumors by means of artificial intelligence," *Med. Phys.*, vol. 19, pp. 1475–1481, 1992.

[20] J. Kilday, F. Palmieri, and M. D. Fox, "Classifying mammographic lesions using computerized image analysis," *IEEE Trans. Med. Imag.*, vol. 12, pp. 664–669, Dec. 1993.

[21] M. F. McNitt-Gray, H. K. Huang, and J. W. Sayre, "Feature selection in the pattern classification problem of digital chest radiograph segmentation," *IEEE Trans. Med. Imag.*, vol. 14, pp. 537–547, Sept. 1995.

[22] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, R. A. Schmidt, and K. Doi, "Automated computerized classification of malignant and benign masses on digitized mammograms," *Acad. Radiol.*, vol. 5, pp. 155–168, 1998.

[23] M. Jordan, and R. A. Jacobs, "Hierarchical mixture of experts and EM algorithm," *Neural Comput.*, vol. 6, pp. 181–214, 1994.

[24] L. Hadjiiski and P. Hopke, "Design of large scale models based on multiple neural network approach," *Intelligent Engineering Systems Through Artificial Neural Networks*. ASME, 1997, vol. 7, pp. 61–66.

[25] S. Grossberg, "Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors," *Biolog. Cybern.*, vol. 23, no. 3, pp. 121–134, 1976.

[26] G. A. Carpenter and S. Grossberg, "ART 2: Self-organization of stable category recognition codes for analog input patterns," *Appl. Opt.*, vol. 26, no. 23, 1, pp. 4919–4930, Dec. 1987.

[27] G. A. Carpenter, S. Grossberg, and D. B. Rosen, "ART 2-A: An adaptive resonance algorithm for rapid category learning and recognition," *Neural Networks*, vol. 4, no. 4, pp. 493–504, 1991.

[28] G. A. Carpenter and S. Grossberg, "Integrating symbolic and neural processing in a self-organizing architeture for pattern recognition and prediction," in *Artificial Intelligence and Neural Networks: Steps toward Principled Integration*. New York: Academic, 1994.

[29] G. A. Carpenter and N. Markuzon, "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," *Neural Networks*, vol. 11, no. 2, pp. 323–336, Mar. 1998.

[30] Y. Xie, P. K. Hopke, and D. Wienke, "Airborne particle classification with a combination of chemical composition and shape index utilizing an adaptive resonance artificial neural network," *Environ. Sci. Technol.*, vol. 28, no. 11, pp. 1921–1928, 1994.

[31] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Texture features for image classification," *IEEE Trans. Syst., Man, Cybern.*, vol. 3, pp. 610–621, Nov. 1973.

[32] M. M. Galloway, "Texture analysis using gray level run length," *Comput. Graph. Image Processing*, vol. 4, pp. 172–179, 1975.

[33] M. J. Norusis, *SPSS Professional Statistics 6.1*. Chicago, IL: SPSS, 1993.

[34] M. M. Tatsuoka, "Multivariate Analysis," *Techniques for Educational and Psychological Research*. New York: Macmillan, 1988.

[35] C. E. Metz, "ROC methodology in radiographic imaging," *Invest. Radiol.*, vol. 21, pp. 720–733, 1986.

[36] C. E. Metz, J. H. Shen, and B. A. Herman, "New methods for estimating a binomial ROC curve from continuously distributed test results," presented at the *1990 Annu. Meeting American Statistical Association*, Anahaim, CA, 1990.

[37] B. Sahiner, H. P. Chan, N. Petrick, R. Wagner, and L. Hadjiiski, "The effect of sample size on feature selection in computer-aided diagnosis," *Proc. SPIE*, vol. 3661, pp. 499–510, 1999.